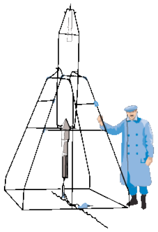
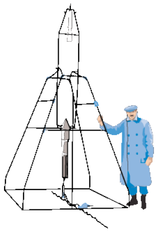


Introduction To NASA High End Computing (HEC) WAN File Accessing Experiments/Demonstrations At SC10

Pat Gary
Pat.Gary@nasa.gov
Computational and Information Sciences and Technology Office (CISTO), Code 606
NASA Goddard Space Flight Center

October 16, 2010

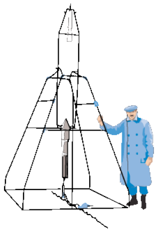
Information Supporting NASA HEC WAN File Accessing Experiments/Demonstrations At SC10



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC10

Topics

- Overall objectives
  - Build on last year
- But with significantly updated test matrix
- And with significantly updated:
  - NASA participants
  - NASA partners
  - Vendor loaned equipment
  - NASA-built net-test workstations
  - LAN & WAN test configurations
- Also supporting other demos, e.g., ANL, DICE, LAC, & NOAA



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC10

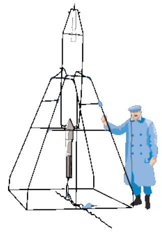
Objectives of NASA HEC WAN File Accessing Experiments

- Determine optimal ‘tuning parameter” settings to obtain maximum user throughput performance with several traditional and new (or emerging) disk-to-disk file-copying utilities when operating over multi-10Gbps WANs using new state-of-the-art high performance workstations and servers

- Inter-compare throughput findings from traditional versus new file-copying utilities

- As a baseline, determine maximum memory-to-memory throughput performance among the workstations and servers using nuttcp (http://www.nuttcp.org/)

- Are an integral part of GSFC/HEC’s 20, 40 & 100 Gbps Network Testbed Plan (http://science.gsfc.nasa.gov/606.1/docs/HECN_10G_Testbeds_082210.pdf)



Optimizing Wide-Area File Transfer for 10-Gbps and Beyond

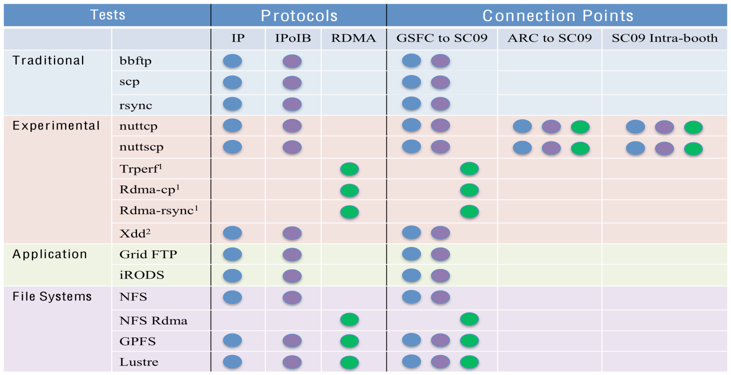
- Demonstrations of network-performance testing, wide-area file systems, and file transfer applications ranging from traditional to experimental were provided in the NASA research exhibit at the SC09 conference, Portland, OR, Nov. 16–19.

- Jointly planned by GSFC's High End Computer Network Team and NCCS' Advanced Development Team, an indication of the wide-area file transfer applications demonstrated and evaluated is shown in the Data Transfer Test Matrix (top figure) and the WAN infrastructure and servers tested are shown in the configuration diagram (bottom figure).

- Demonstration highlights included over 100 gigabits per second (Gbps) uni-directional memory-to-memory data transmissions between in-booth servers, 40-Gbps bi-directional memory-to-memory data transmissions between servers in-booth and at ARC, 10-Gbps disk-to-disk data transfers between in-booth servers, between servers in-booth and at ARC, and between servers in-booth and at GSFC.

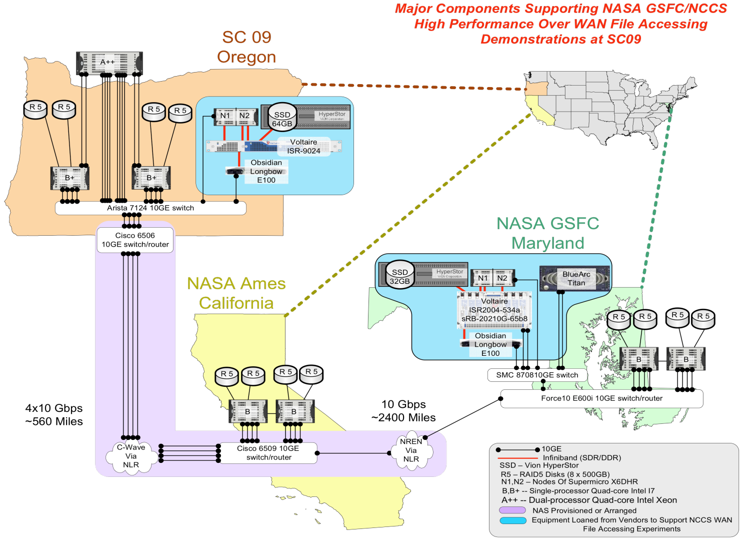
POC: Pat Gary, Pat.Gary@nasa.gov,
(301) 286-9539, GSFC Computational and
Information Sciences and Technology Office

High Performance Wide Area Data Transfer Test Matrix



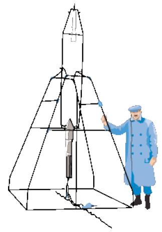
Tests		Protocols			Connection Points		
		IP	IPoIB	RDMA	GSFC to SC09	ARC to SC09	SC09 Intra-booth
Traditional	bbftp	🔵	🟣		🔵 🟣		
	scp	🔵	🟣		🔵 🟣		
	rsync	🔵	🟣		🔵 🟣		
Experimental	nuttcp	🔵	🟣		🔵 🟣	🔵 🟣 🟢	🔵 🟣 🟢
	nuttscp	🔵	🟣		🔵 🟣	🔵 🟣 🟢	🔵 🟣 🟢
	Trperf[1]			🟢	🟢		
	Rdma-cp[1]			🟢	🟢		
	Rdma-rsync[1]			🟢	🟢		
	Xdd[2]	🔵	🟣		🔵 🟣		
Application	Grid FTP	🔵	🟣		🔵 🟣		
	iRODS	🔵	🟣		🔵 🟣		
File Systems	NFS	🔵	🟣		🔵 🟣		
	NFS Rdma			🟢	🟢		
	GPFS	🔵	🟣	🟢	🔵 🟣 🟢		
	Lustre	🔵	🟣	🟢	🔵 🟣 🟢		

[1] Courtesy of Obsidian Research.
[2] End-to-end file transfers supported by the Oak Ridge National Laboratory Extreme Scale System Center and the Department of Defense.



Major Components Supporting NASA GSFC/NCCS High Performance Over WAN File Accessing Demonstrations at SC09

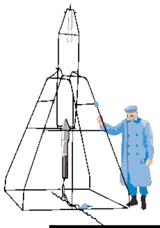
Figures: Data Transfer Test Matrix (Top) and WAN infrastructure and servers tested (bottom) during SC09.



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC10

Reference Articles & Websites Per SC09 Demos

- Introduction To NASA High End Computing (HEC) WAN File Accessing Experiments/Demonstrations At SC09
  - http://science.gsfc.nasa.gov/606.1/docs/SC09_NCCS-demos_mini2_021910.pdf
- "Optimizing Wide-Area File Transfers for 10 Gbps and Beyond”
  - http://www.nas.nasa.gov/SC09/PDF/Datasheets/Gary_OptimizingWide.pdf
- “NASA Successfully Demonstrates Remote High-speed Encrypted InfiniBand Applications Over National LambdaRail”
  - http://www.virtualpressoffice.com/detail.do?contentId=208703&companyId=3273&showId=1215381715818
- “NASA Demos Secure Coast-to-Coast Backup at Full Wire Speed Using Obsidian's New Longbow E100 and DSYNC”
  - http://www.virtualpressoffice.com/publicsiteContentFileAccess?fileContentId=206528&fromOtherPageToDisableHistory=Y&menuName=News&sId=1215381715818&sInfo=Y
- NASA use of NLR during SC09
  - http://www.flickr.com/photos/nationallambdarail/4189002873/



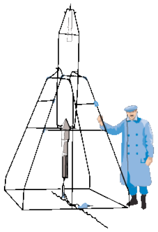
Test Matrix for Optimizing Wide-Area File Transfer

Tool	Type
aspera fasp**	Disk ↔ Disk
nuttscp	Disk ↔ Disk
GridFTP	Disk ↔ Disk
RocketStream**	Disk ↔ Disk
iRODS	Disk ↔ Disk
bbFTP	Disk ↔ Disk
FDT	Disk ↔ Disk
HPN-SCP	Disk ↔ Disk
dsync*	Disk ↔ Disk
xdd	Disk ↔ Disk
rsync	Disk ↔ Disk
FTP	Disk ↔ Disk
gpfs	Mem/Disk ↔ Disk
pNFS	Mem/Disk ↔ Disk
NFS	Mem/Disk ↔ Disk
NFS-RDMA*	Mem/Disk ↔ Disk
nuttcp	Mem ↔ Mem

*RDMA options – IB/Obsidian, iWARP, RoCE

**Commercial product

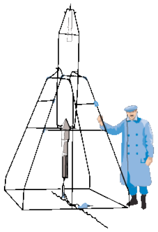
- Desired measurements: disk-to-disk file-copying-throughput performance (in Gbps), plotted against different file-sizes and different conditions
- Key single-file-sizes in GBs: 16, 32, 64, 128
- Primary different conditions:
  - File-copying-applications, e.g., GridFTP, bbFTP, nuttscp, ...
    - Both well-established and experimental/emerging ones
  - Key round trip times (RTTs) in milliseconds: 0, 15, 90, 180
    - Corresponding very roughly to LAN, large MAN/RON, trans-USA/WAN, trans-Atlantic
- Secondary different conditions, when time permits:
  - Several-file-sizes yet with a constant 256 GB total volume: 16@16, 8@32, 4@64, 2@128
  - Many-file-sizes yet with a constant total volume
  - Cases with packet loss, corruption, etc
  - Real tests
    - @10Gbps: GSFC-GSFC (RTT=0); GSFC-StarLight (RTT=~17); GSFC-ARC (RTT=~87)
    - @20Gbps: GSFC-GSFC; GSFC-StarLight
    - @40Gbps: GSFC-GSFC; GSFC-SC10 (RTT=~35)



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC10

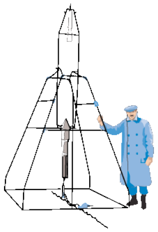
Elaboration of Test Matrix for Optimizing Large File Transfers Over Wide Areas (1 of 2)

- Desired measurements (y-axis): disk-to-disk file-copying-throughput performance (in Gbps), plotted against different file-sizes and different conditions
- Key single-file-sizes in GBs: 16, 32, 64, 128
- Primary different conditions:
  - File-copying-applications, e.g., GridFTP, bbFTP, nuttscp, …
    - Both well-established and experimental/emerging ones
  - Key round trip times (RTTs) in milliseconds: 0, 15, 90, 180
    - Corresponding very roughly to LAN, large MAN/RON, trans-USA/WAN, trans-Atlantic



Elaboration of Test Matrix for Optimizing Large File Transfers Over Wide Areas (2 of 2)

- Secondary different conditions, when time permits:
  - Several-file-sizes yet with a constant 256 GB total volume: 16@16, 8@32, 4@64, 2@128
  - Many-file-sizes yet with a constant total volume, e.g., a single directory of small-to-medium-to-partly-large size files
  - Cases with packet loss, corruption, etc, particularly if they can be controlled by netem and "be realistic"
  - Real tests near-term
    - @10Gbps: GSFC-GSFC (RTT=0); GSFC-StarLight (RTT=~17); GSFC-ARC (RTT=~87)
    - @20Gbps: GSFC-GSFC; GSFC-StarLight
  - Real tests very soon
    - @40Gbps: GSFC-GSFC; GSFC-SC10 (RTT=~35)



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC10

NASA HEC WAN File Accessing Team

- GSFC NASA Center for Climate Simulation (NCCS)
  - Dan Duffy/GSFC

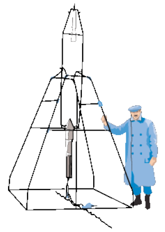
  - Hoot Thompson/PTP
  - Kirk Hunter/PTP

- GSFC/NCCS HEC Network (HECN) Team
  - Pat Gary/GSFC
  - Bill Fink/GSFC

  - Paul Lang/ADNET
  - Jeff Martz/ADNET
  - Aruna Muppalla/ADNET
  - Mike Stefanelli/ADNET

- ARC/CIO Network Team
  - Kevin Jones/ARC
  - Hugh LaMaster/ARC

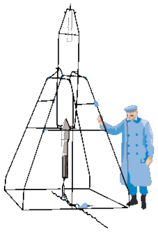
  - Dave Hartzel/CSC
  - Mark Foster/CSC
  - Matt Mountz/CSC



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC10

NASA Partners in “Using 100G Network Technology in Support of Petascale Science” Special Demos (1 of 2)

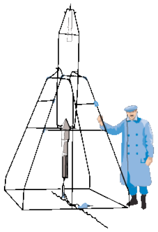
- International Center for Advanced Internet Research (iCAIR), PI: Dr. Joe Mambretti/Northwestern University
- Laboratory for Advanced Computing (LAC), PI: Dr. Bob Grossman/UIC
- Mid-Atlantic Crossroads (MAX), PM: Peter O’Neil/UMCP
- National LambdaRail (NLR), POC: Bonnie Hurst/NLR
- National Oceanic and Atmospheric Administration (NOAA), POC: Jerry Janssen
- SCinet Research Sandbox (SRS), Chair: Rodney Wilson/Ciena
- Vendors who loaned equipment – see following charts



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC10

Acknowledgement of Vendor Equipment On Loan

- Ciena: Two Optical Multiservice Edge 6500 units each with 100G transport and 10x10G-to-1x100G muxponder interfaces
- Cisco: Two CRS-3 switch/routers each with 100GE and 14x10GE interfaces, plus use of a third CRS-3 with two 100GE interfaces
- ColorChip: Six DragonFly 40G-LR (up to 10km) QSFP transceivers
- Extreme Networks: Two VIM3-40G4X 4-port 40GE modules (for Summit X650 10GE switches)
- Fusion-io: Two Octal cards (SSDs on PCIe Gen2 x16)
- HP: Two ProLiant DL580 G7 servers with two 2x10GE NICs and one QDR IB HCA
- Panduit: Two CN1 Net-Access Switch Cabinets



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC10

Acknowledgement of Partner Contributions (Partial)

- iCAIR: Many 10GE connections in and through the StarLight@Chicago
- LAC: Exhibit booth use at SC10
- MAX: 4x10G lambdas between GSFC@Greenbelt & NLR@McLean/DC
- NLR: 4x10G lambdas between DC & StarLight and another 8x10G pathways (two are really Cisco C-Waves; one is dedicated for NASA) between StarLight & Baton Rouge (plus coordination across the Louisiana Optical Network Initiative (LONI) regional optical network (RON) to SC10@NewOrleans)
- NOAA: Exhibit booth use at SC10
- SRS: Cost-discounted fiber-pair-bundles between the exhibit booths of NASA, NCDM-LAC/iCAIR, NOAA & SCinet NOC

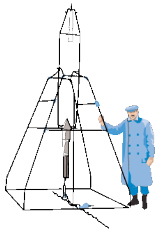
Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC10

NASA Partners in "Using 100G Network Technology in Support of Petascale Science" Special Demos (2 of 2)

- Internet2, POC: Chris Robb/Internet2

Acknowledgement of Partner Contributions (Partial)

- Internet2: Use of their 1x100G pathway between StarLight@Chicago & SC10@NewOrleans for their Multi-Vendor 100GigE Demo Between Chicago and SC10



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC10

Identification of GSFC/HECN Equipment Used

- A, A+ net-test workstations*: see following pages
- Arista: 7124S & 7148SX 10GE switches
- B, B+ net-test workstations*: see following pages
- C net-test workstations*: see following pages
- Extreme Networks: Summit X650 10GE switch
- Force10: E600i & S2410 10GE switches
- X++ net-test workstations*: see following pages
- XSSD, XSSD++ net-test workstations*: see following pages

*Advanced hint: A, B & C's are Intel Core i7-based; X & XSSD's are Intel Xeon-based

100 Gigabits per Second Transmissions Achieved Via A Single Workstation

- As part of plans to assess the throughput performance of wide-area file transfer applications, GSFC's High End Computer Network (HECN) Team specified and assembled a workstation that costs less than $11,000 and is capable of over 100 gigabits per second (Gbps) data transmission – 10 times the transmission speed of most high end computers.

- The workstation consists of a 3.2-GHz dual-processor (quad core) Intel Xeon W5580 (Nehalem) with six Myricom dual-port 10-Gigabit Ethernet network interface cards in the PCIe Gen2 slots of a Supermicro X8DAH+-F motherboard.

- Over 100-Gbps aggregate-throughput transmissions from the Xeon-workstation to two Intel Core i7 workstations (also specified and assembled by the HECN Team) were measured using the nuttcp (www.nuttcp.net) network-performance testing tool.

- Demonstrations of these workstations supporting network-performance testing, wide-area file systems, and file transfer applications ranging from traditional to experimental are planned in the NASA research exhibit at the SC09 conference, Portland, OR, Nov. 16–19 .

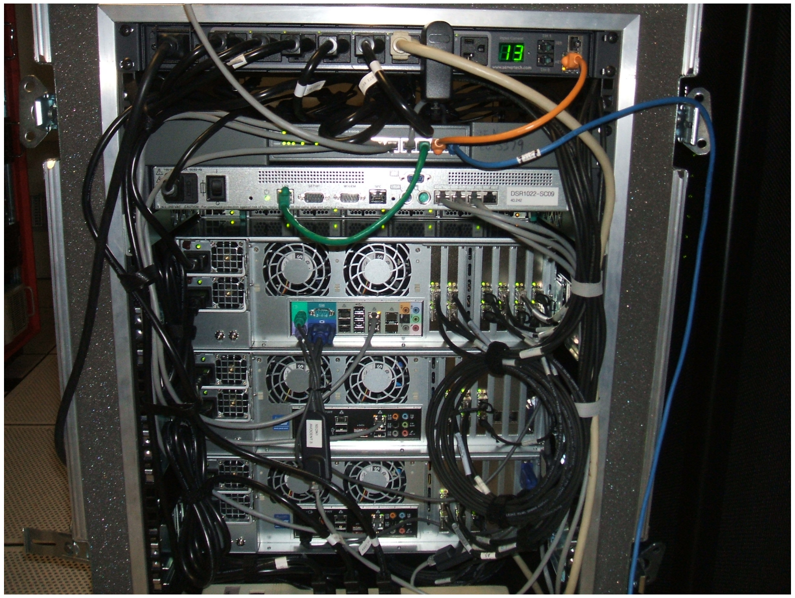


Figure: Xeon and two Core i7 workstations (bottom) interconnected with 10 Gigabit Ethernet switch and management units (top) in a rack for shipping to SC09.

POC: Bill Fink, William.E.Fink@nasa.gov, (301) 286-7924, GSFC Computational and Information Sciences and Technology Office

10-Gbps Disk-to-Disk File Copies Achieved Via Workstations Costing Less Than $7,000

- As part of plans to assess the throughput performance of wide-area file transfer applications, GSFC's High End Computer Network Team specified and assembled workstations that individually costs less than $7,000 and are capable of over 10 gigabits per second (Gbps) disk-to-disk file copying.
- Each workstation consists of a 3.2-GHz single-processor (quad core) Intel Core i7 (Nehalem) with two HighPoint RocketRaid 4320 RAID disk controllers and a Myricom 10 Gigabit Ethernet network interface card in the PCIe Gen2 slots of a Asus P6T6 WS Revolution motherboard. Each RAID controller hosts eight Western Digital WD5001AALS 500-gigabyte disks.
- Over 10-Gbps disk-to-disk file-copying throughput between two of the workstations was measured using the nuttscp (www.nuttcp.net) file copying tool.
- Demonstrations of these workstations supporting network-performance testing, wide-area file systems, and file transfer applications ranging from traditional to experimental are planned in the NASA research exhibit at the SC09 conference, Portland, OR, November 16–19 .

Figure: Two Core i7 workstations interconnected via 10 Gigabit Ethernet in test configuration prior to shipping to SC09.

POC: Bill Fink, William.E.Fink@nasa.gov, (301) 286-7924, GSFC Computational and Information Sciences and Technology Office

17.8-Gbps Disk-to-Disk File Copies Achieved Via Workstations Costing Less Than $9,000

- As part of plans to assess the throughput performance of wide-area file transfer applications, GSFC's High End Computer Network Team specified and assembled workstations that individually costs less than $9,000 and are capable of over 17.8 gigabits per second (Gbps) disk-to-disk file copying.

- Each workstation consists of a 3.2-GHz single-processor (quad core) Intel Core i7 (Nehalem) with four HighPoint RocketRaid 4320 RAID disk controllers and a Myricom 2-port 10 Gigabit Ethernet network interface card in the PCIe Gen2 slots of a Asus P6T6 WS Revolution motherboard. Each RAID controller hosts eight Western Digital WD5001AALS 500-gigabyte disks.

- Over 17.8-Gbps disk-to-disk file-copying throughput between two of the workstations was measured using the nuttscp (www.nuttcp.net) file copying tool.

- While SSD technology is next to be investigated, parallelism of multiple cores and multiple streams is likely to be key to going to 40-Gbps and beyond, since individual cores are not getting significantly faster.

Figure: Right case houses Core i7 cores, DDR3 memory, NIC, two "internal" controllers each with eight disks and two "external" controllers; left case houses sixteen SAS-connected disks.

POC: Bill Fink, William.E.Fink@nasa.gov, (301) 286-7924, GSFC Computational and Information Sciences and Technology Office

Aggregate 55+ Gigabits per Second (Gbps) Transmits, 52+ Gbps Receives and 75+ Gbps Bi-Directional Transmissions Achieved Via A Single Workstation With a Single 6x10-Gigabit Ethernet Network Interface Card

- As part of plans to assess the throughput performance of wide-area file transfer applications, GSFC's High End Computer Network (HECN) Team tested a HotLava six-port 10-Gigabit Ethernet network interface card (NIC) in a HECN Team-assembled workstation that costs less than $ 6,800 with the NIC and achieved aggregate 55+ Gbps transmits, 52+ Gbps receives and 75+ Gbps bi-directional memory-to-memory data transmissions.

- The workstation consists of a 3.2-GHz single-processor (quad core) Intel Core i7 (Nehalem) with one HotLava NIC in one PCIe Gen2 x16 slot of an Asus P6T6 WS Revolution motherboard.

- Transmissions between the above workstation and two other HECN Team-assembled Intel Core i7 workstations with other NICs were measured using the nuttcp (www.nuttcp.net) network-performance testing tool.

- Demonstrations of similar workstations supporting 100 Gbps network testing and near-40 Gbps file transfer applications are planned in the NASA research exhibit at the SC10 conference, New Orleans, LA, Nov. 15–18.

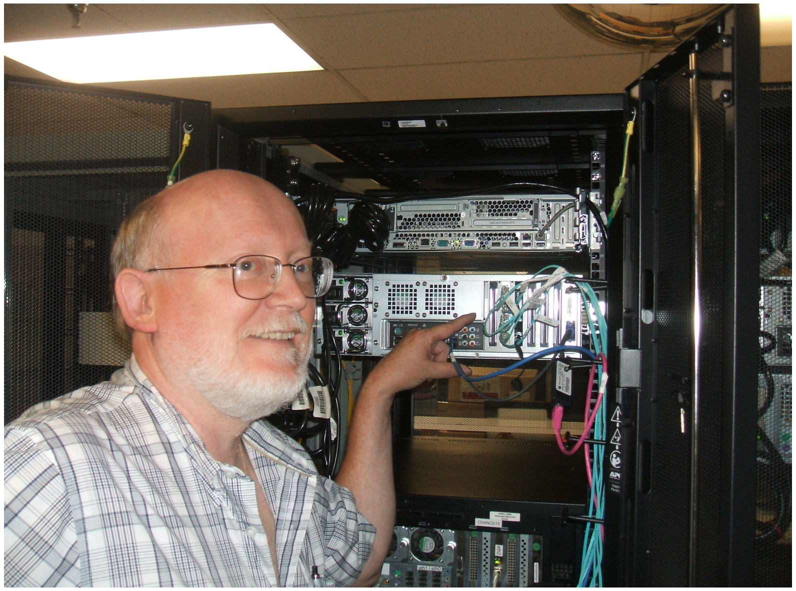
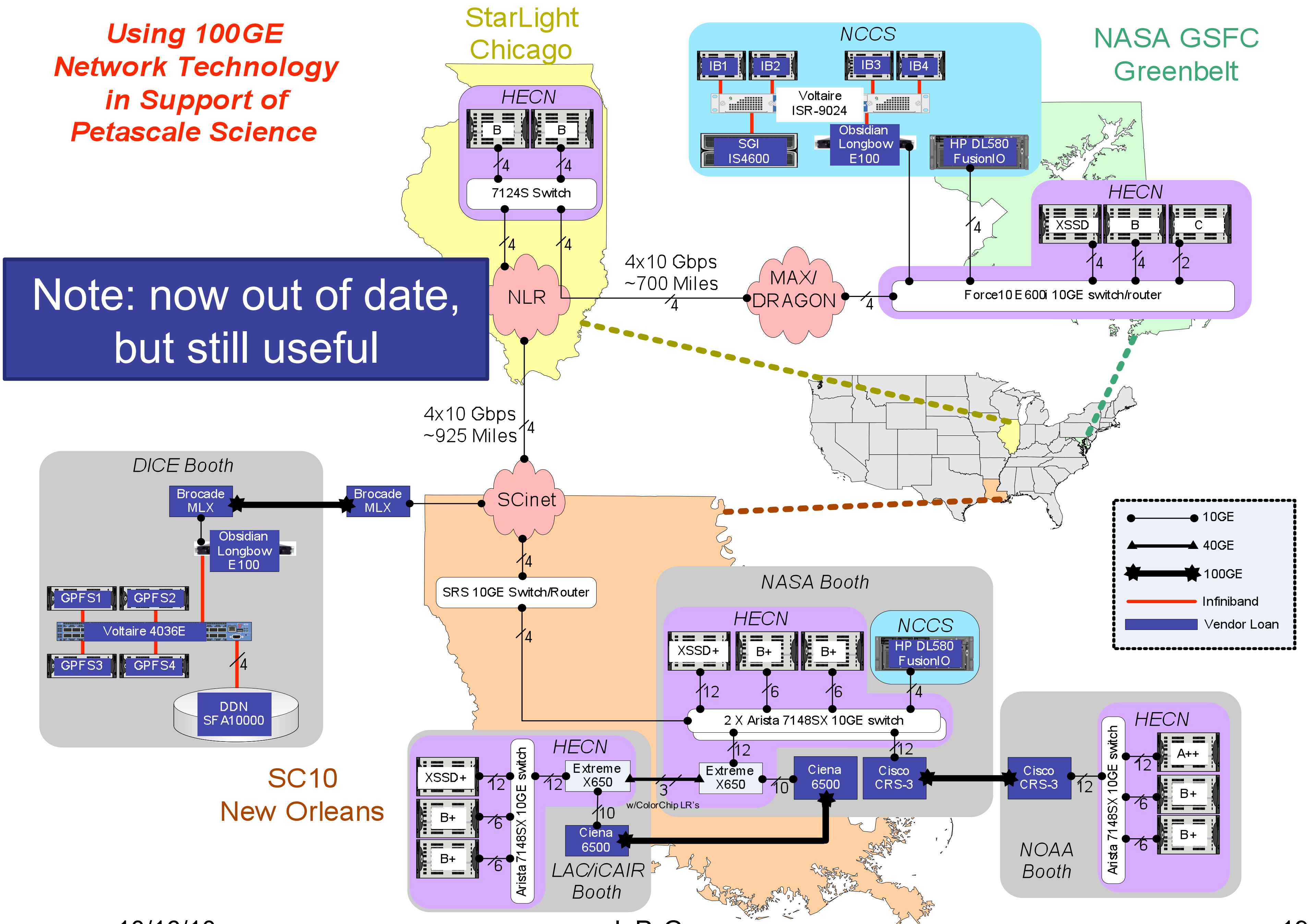


Figure: Bill Fink, author of nuttcp and the throughput performance tests, pointing to the 6x10GE HotLava NIC in the HECN Team's Intel Core i7 based workstation.

POC: Bill Fink, Bill.Fink@nasa.gov, GSFC Computational and Information Sciences and Technology Office

Using 100GE Network Technology in Support of Petascale Science

Note: now out of date, but still useful



StarLight Chicago

HECN

B B

4 4

7124S Switch

4 4

NLR

4x10 Gbps ~700 Miles

4

MAX/ DRAGON

4

NCCS

IB1 IB2 IB3 IB4

Voltaire ISR-9024

SGI IS4600

Obsidian Longbow E100

HP DL580 FusionIO

NASA GSFC Greenbelt

4

HECN

XSSD B C

4 4 2

Force10 E600i 10GE switch/router

4x10 Gbps ~925 Miles

4

DICE Booth

Brocade MLX

Obsidian Longbow E100

GPFS1 GPFS2

Voltaire 4036E

GPFS3 GPFS4

4

DDN SFA10000

Brocade MLX

SCinet

4

SRS 10GE Switch/Router

4

NASA Booth

HECN

XSSD+ B+ B+

12 6 6

NCCS

HP DL580 FusionIO

4

2 X Arista 7148SX 10GE switch

12 12

Extreme X650

10

Ciena 6500

Cisco CRS-3

3

w/ColorChip LR's

HECN

XSSD+ B+ B+

12 6 6

Arista 7148SX 10GE switch

12

Extreme X650

10

Ciena 6500

LAC/iCAIR Booth

Cisco CRS-3

12

HECN

Arista 7148SX 10GE switch

12 6 6

A++ B+ B+

NOAA Booth

SC10 New Orleans

●—●	10GE
▲—▲	40GE
★━★	100GE
🟥	Infiniband
🟦	Vendor Loan

National Aeronautics and Space Administration

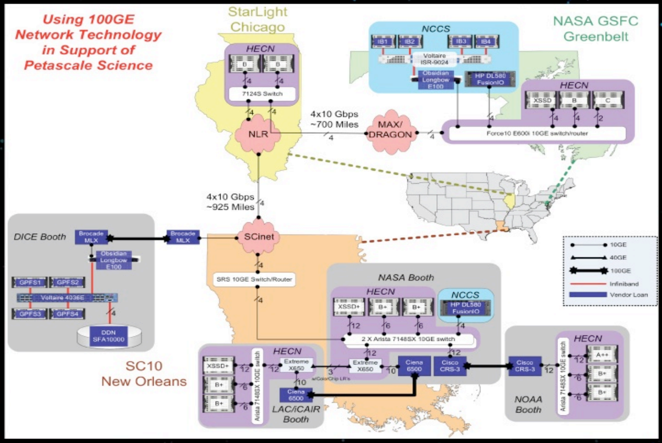
Network Testbed for Enhanced Earth Science Simulations

NASA Earth science models are generating more than 120 terabytes per month in wide-area network transfers. With geographically distributed systems, it is vital to provide highly optimized, high-end network services to store, share, and analyze large amounts of data.

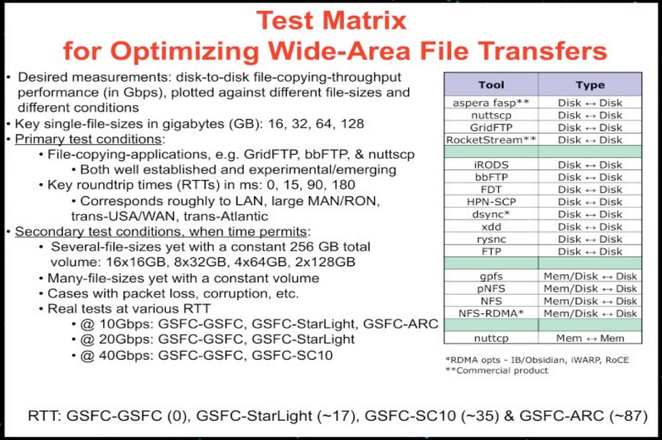
To improve data movement and management for this research, a testbed of advanced hardware and software technologies has been installed between the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center and the StarLight facility in Chicago. During SC10, this testbed has been extended to the exhibits of the University of Illinois at Chicago's Laboratory for Advanced Computing, the National Oceanic and Atmospheric Administration, and NASA.

- The testbed utilizes 10-gigabit-per-second through 100-gigabit-per-second technologies including national research and development (R&D) networks
- Wide-area issues are addressed through a suite of tests ranging from experimental wire-speed tests, traditional and emerging file transfer applications, and file systems

The successes of this testbed will allow scientists, particularly those engaged in climate research sponsored by NASA's Science Mission Directorate, to focus on their missions rather than on day-to-day tasks associated with data management.



Major components and partners supporting the NASA Center for Climate Simulation's high-performance file transfers over wide-area network testbeds and different approaches to 100 gigabit-per-second networking. *Pat Gary, NASA/Goddard*



Test Matrix for Optimizing Wide-Area File Transfers

- Desired measurements: disk-to-disk file-copying-throughput performance (in Gbps), plotted against different file-sizes and different conditions
- Key single-file-sizes in gigabytes (GB): 16, 32, 64, 128
- Primary test conditions:
  - File-copying-applications, e.g. GridFTP, bbFTP, & nuttscp
    - Both well established and experimental/emerging
  - Key roundtrip times (RTTs) in ms: 0, 15, 90, 180
    - Corresponds roughly to LAN, large MAN/RON, trans-USA/WAN, trans-Atlantic
- Secondary test conditions, when time permits:
  - Several-file-sizes yet with a constant 256 GB total volume: 16x16GB, 8x32GB, 4x64GB, 2x128GB
  - Many-file-sizes yet with a constant volume
  - Cases with packet loss, corruption, etc.
  - Real tests at various RTT
    - @ 10Gbps: GSFC-GSFC, GSFC-StarLight, GSFC-ARC
    - @ 20Gbps: GSFC-GSFC, GSFC-StarLight
    - @ 40Gbps: GSFC-GSFC, GSFC-SC10

Tool	Type
aspera fasp**	Disk ↔ Disk
nuttscp	Disk ↔ Disk
GridFTP	Disk ↔ Disk
RocketStream**	Disk ↔ Disk
iRODS	Disk ↔ Disk
bbFTP	Disk ↔ Disk
FDT	Disk ↔ Disk
HPN-SCP	Disk ↔ Disk
dsync*	Disk ↔ Disk
xdd	Disk ↔ Disk
rysnc	Disk ↔ Disk
FTP	Disk ↔ Disk
gpfs	Mem/Disk ↔ Disk
pNFS	Mem/Disk ↔ Disk
NFS	Mem/Disk ↔ Disk
NFS-RDMA*	Mem/Disk ↔ Disk
nuttcp	Mem ↔ Mem

*RDMA opts - IB/Obsidian, iWARP, RoCE
**Commercial product

RTT: GSFC-GSFC (0), GSFC-StarLight (~17), GSFC-SC10 (~35) & GSFC-ARC (~87)

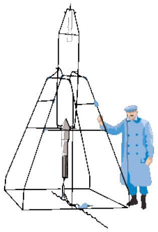
Testing matrix of experimental wire-speed tests, traditional file transfer applications, emerging file transfer applications, and file systems. *Pat Gary, NASA/Goddard*

Pat Gary, NASA Goddard Space Flight Center

SUPERCOMPUTING

www.nasa.gov

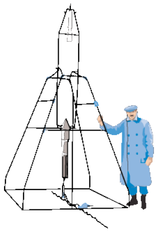
SCIENCE MISSION DIRECTORATE



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC10

Special SC10 Demonstration/Evaluation Experiments

- NASA in collaboration with a set of partners will be conducting a set of individual experiments and demonstrations that collectively are titled "**Using 100G Network Technology in Support of Petascale Science**". The partners include the iCAIR, Internet2, LAC, MAX, NLR, NOAA and SRS as well as the vendors Ciena, Cisco, ColorChip, Extreme Networks, Fusion-io and HP who most generously are willing to allow some of their leading edge 40G/100G network and file server technologies to be involved. The experiments and demonstrations will feature different approaches to 100G networking across the SRS infrastructure between the NCDM-LAC/iCAIR, NASA and NOAA exhibit booths load-stressed by sets of NASA/HECN-built, relatively inexpensive, net-test-workstations that are capable of demonstrating >100Gbps uni-directional nuttcp-enabled memory-to-memory data flows, 80-Gbps aggregate-bidirectional memory-to-memory data transfers, and near 40-Gbps uni-directional disk-to-disk data copies as well as ones between SC10 and StarLight+GSFC across 8x10Gbps network pathways enabled by the NLR and a 1x100 network pathway enabled by the Internet2.



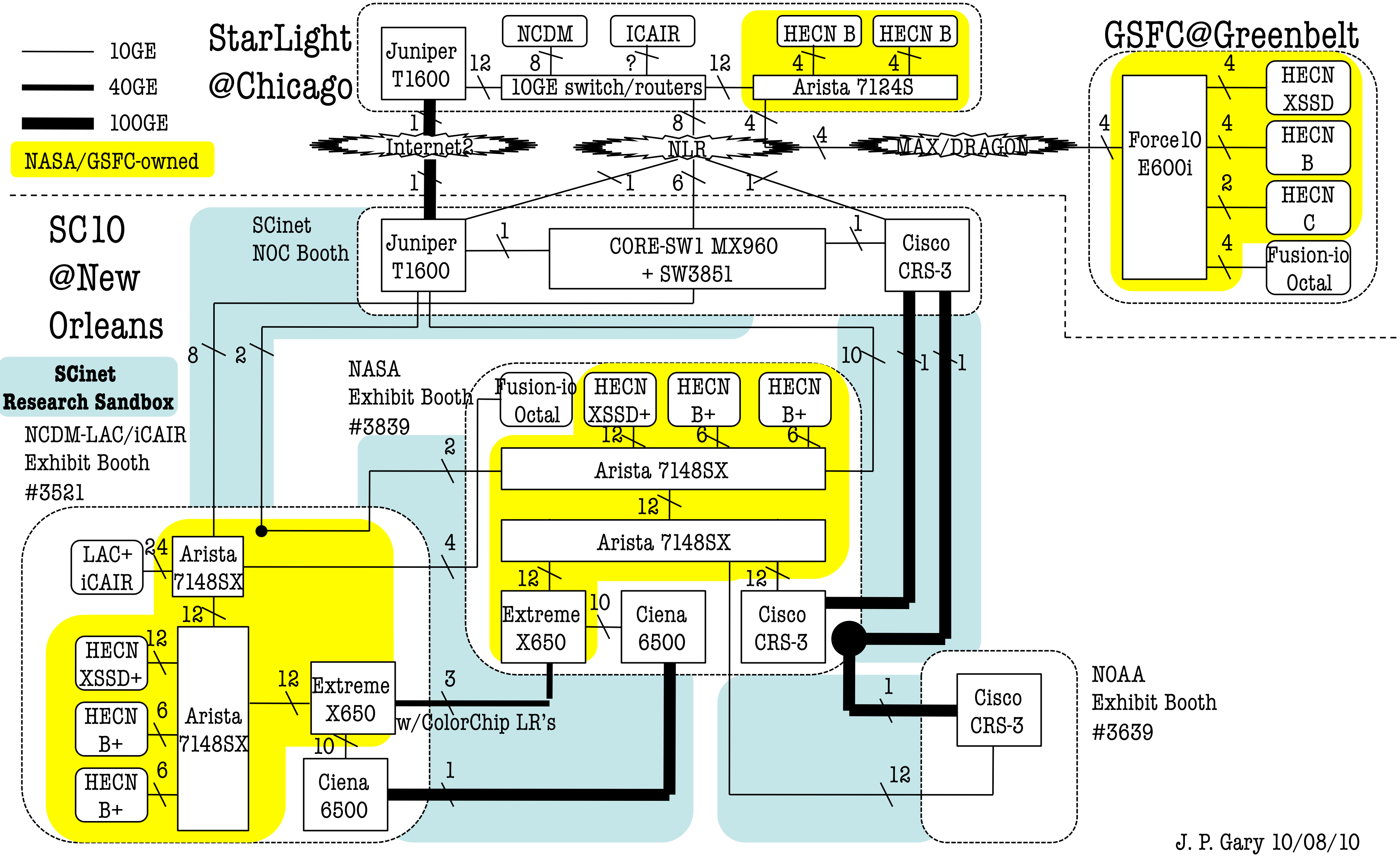
Introduction To
NASA HEC WAN File Accessing
Experiments/Demonstrations At SC10

Special SC10 Demonstration/Evaluation Experiments

- Use a set of the NASA/HECN Team’s network-testing-workstations deployed into each of the NCDM-LAC/iCAIR, NASA and NOAA Exhibit Booths, capable of:
  - >100G uni-directional memory-to-memory data flows
  - >80G aggregate-bidirectional memory-to-memory data flows
  - ~40G uni-directional disk-to-disk file copies (using SSDs)
- Demonstrate/evaluate different vendor-provided 40G/100G network technology solutions with 100G LAN data flows across SCinet Research Sandbox inter-booth fiber
- Use existing 4x10G dedicated pathway across NLR and MAX/DRAGON between GSFC and StarLight, plus a mix of 8 other 10G NLR+Cisco-provisioned pathways and a 1x100G Internet2-provisioned pathway between StarLight and SC10, to conduct science-oriented WAN data flow demonstrations

Using 100G Network Technology in Support of Petascale Science

A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between Chicago and SC10

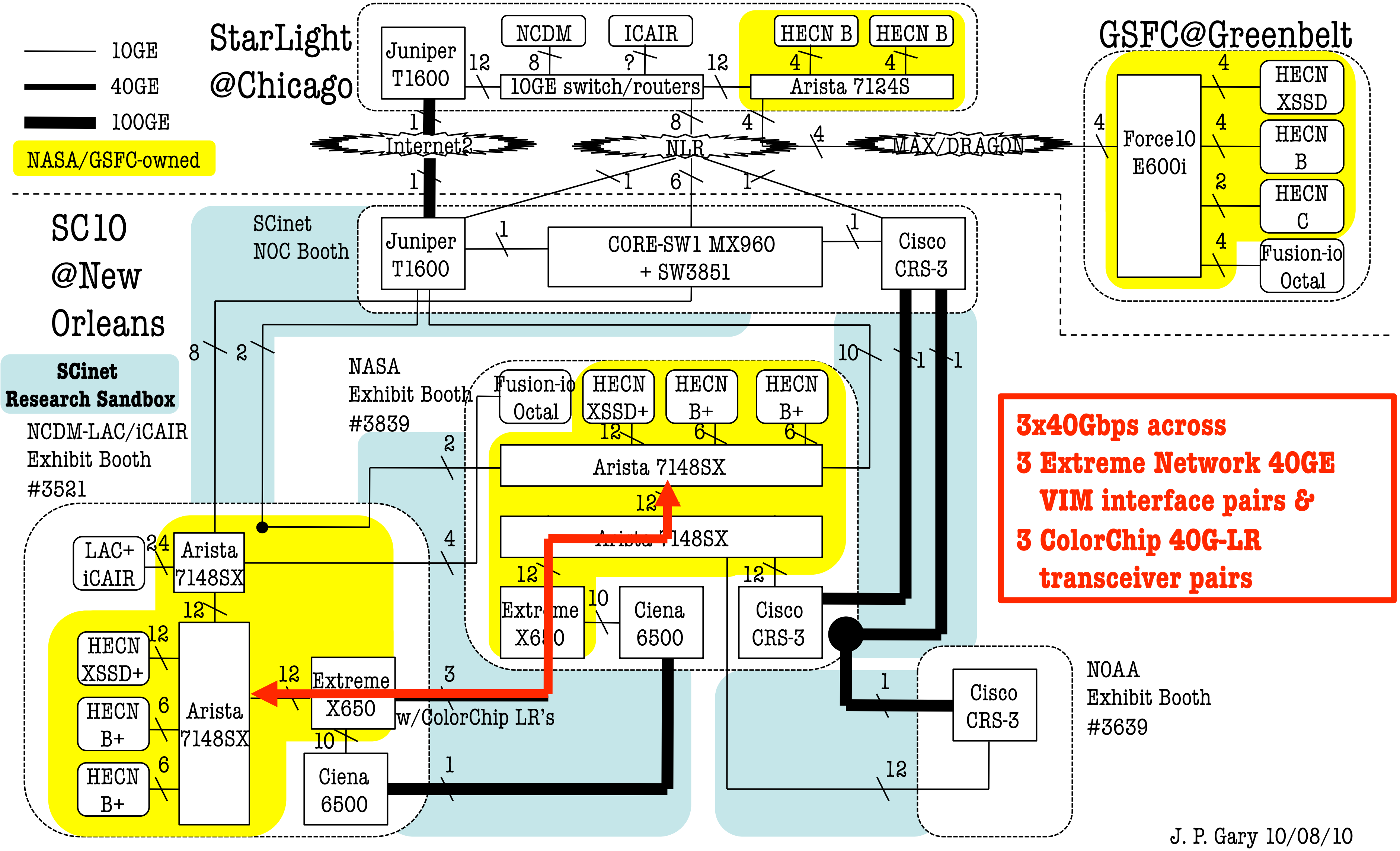


J. P. Gary 10/08/10

Using 100G Network Technology in Support of Petascale Science

A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC

Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between Chicago and SC10

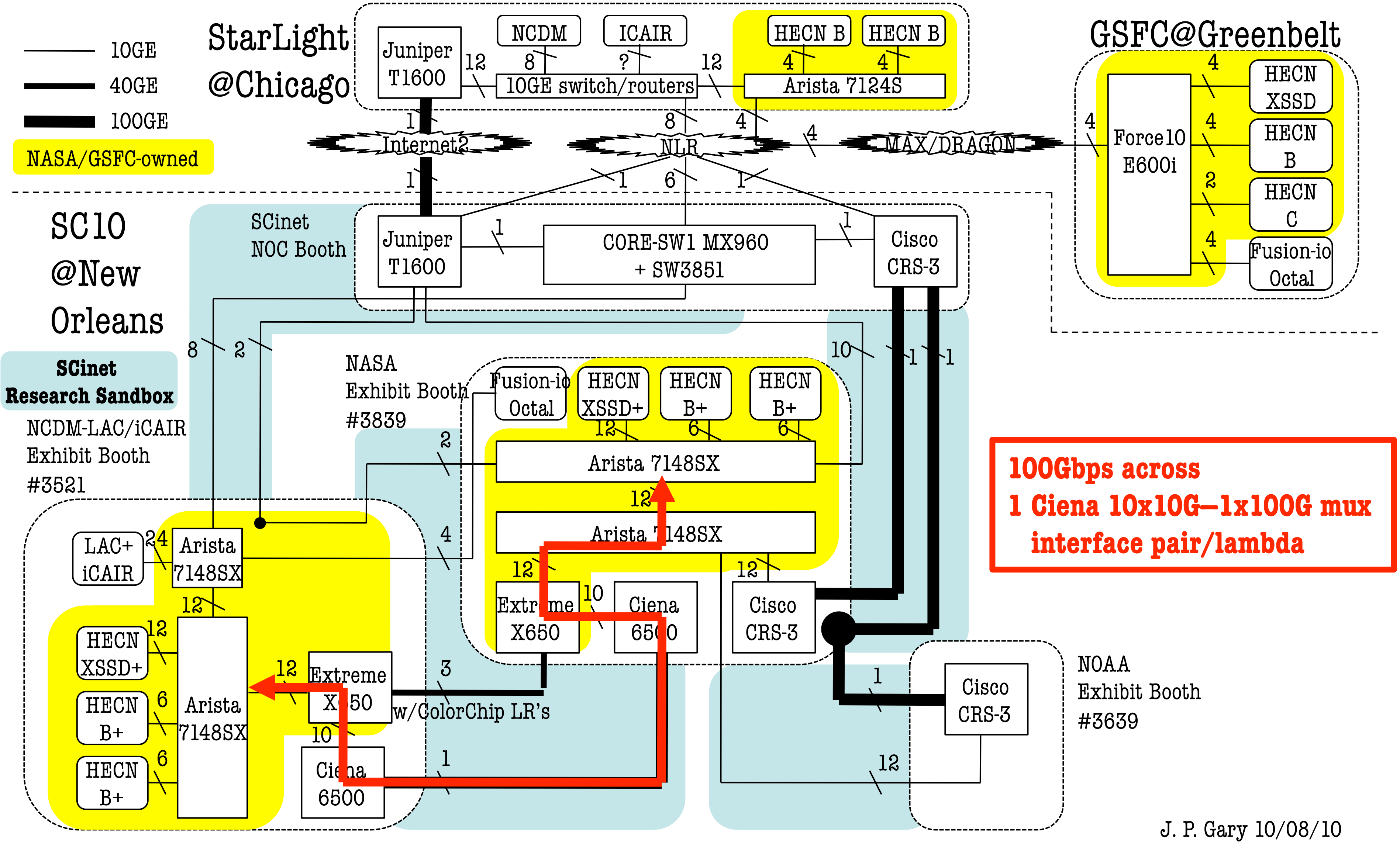


3x40Gbps across
3 Extreme Network 40GE VIM interface pairs &
3 ColorChip 40G-LR transceiver pairs

J. P. Gary 10/08/10

Using 100G Network Technology in Support of Petascale Science

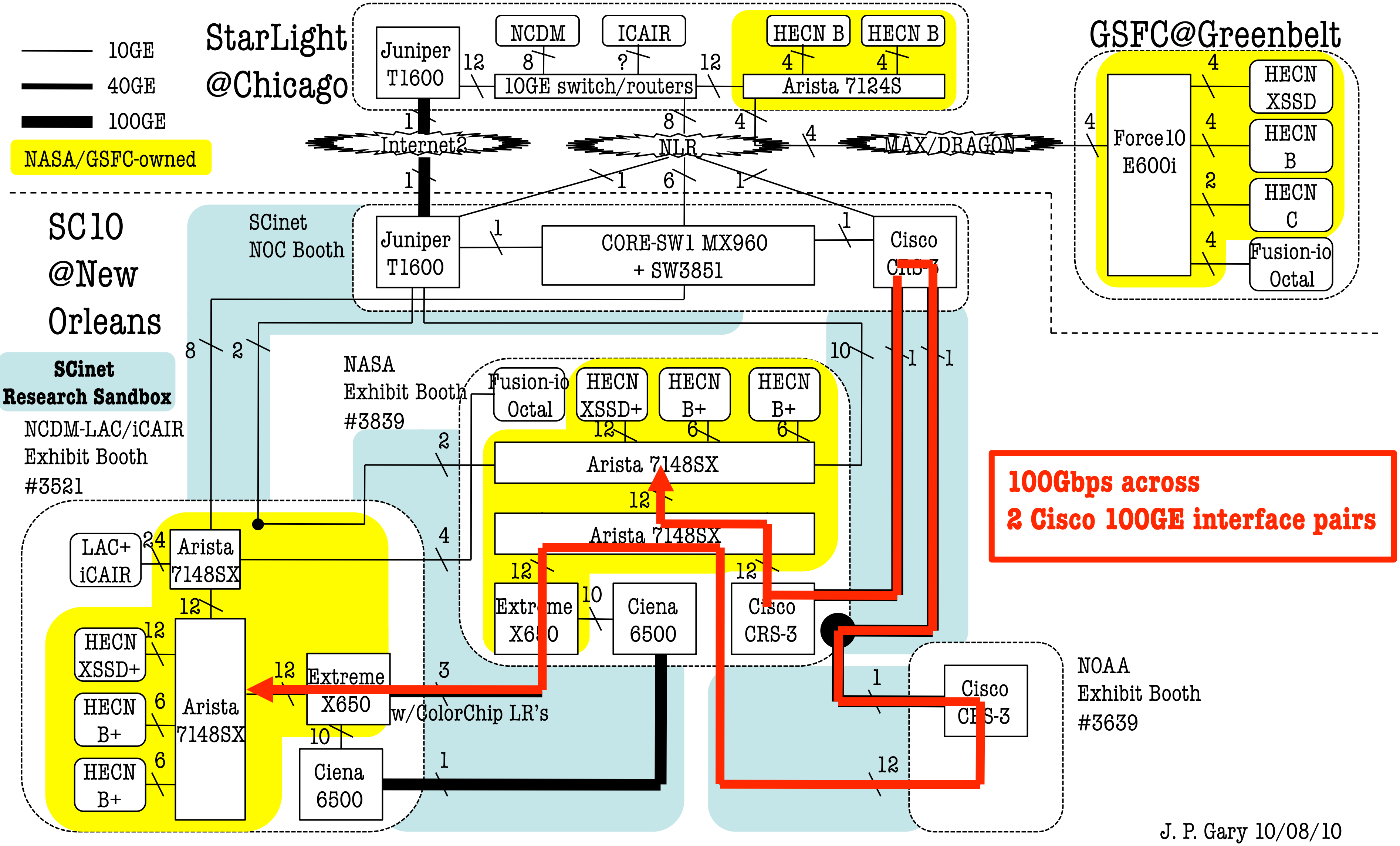
A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between Chicago and SC10



100Gbps across
1 Ciena 10x10G—1x100G mux
interface pair/lambda

Using 100G Network Technology in Support of Petascale Science

A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between Chicago and SC10

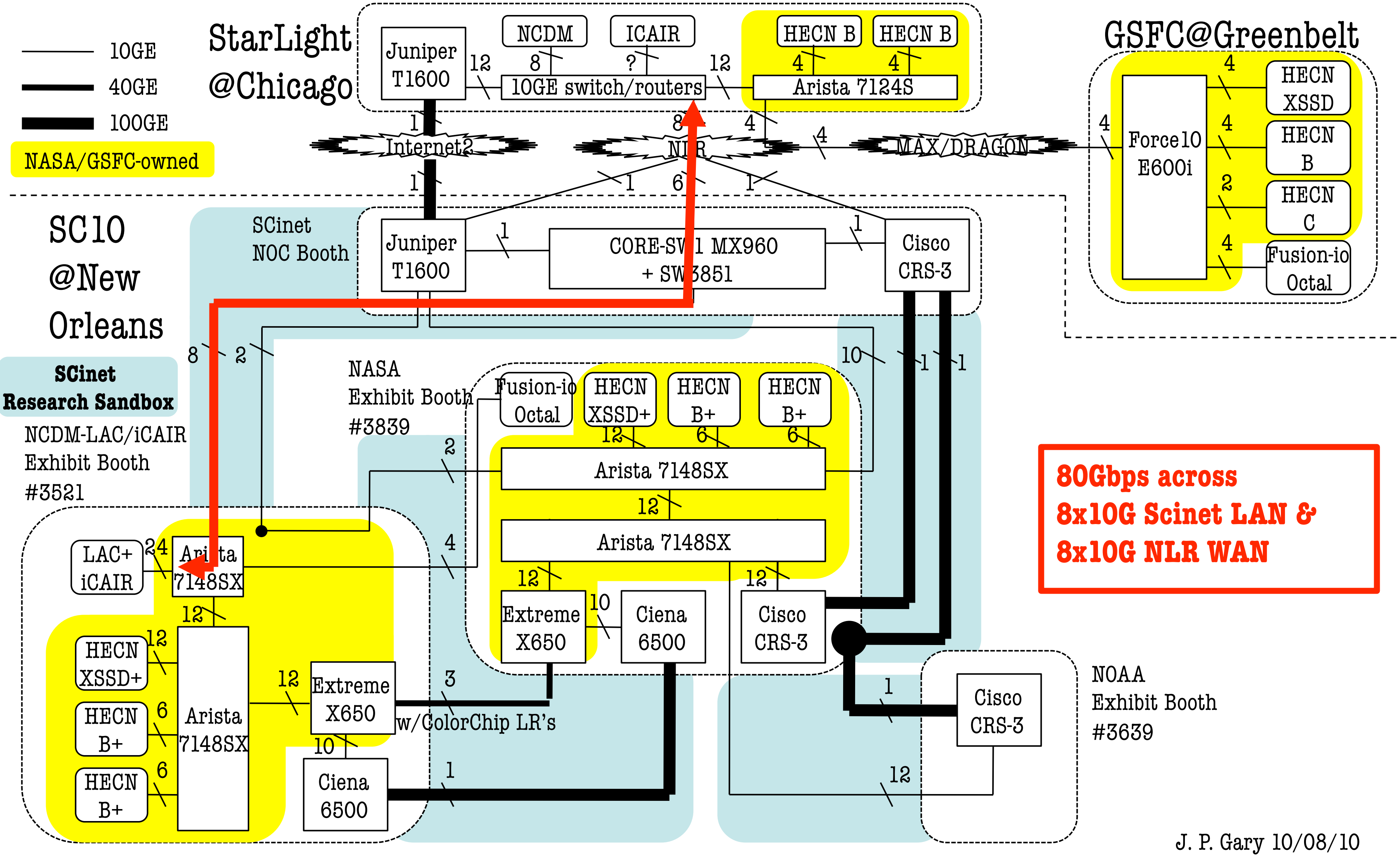


100Gbps across
2 Cisco 100GE interface pairs

J. P. Gary 10/08/10

Using 100G Network Technology in Support of Petascale Science

A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between Chicago and SC10

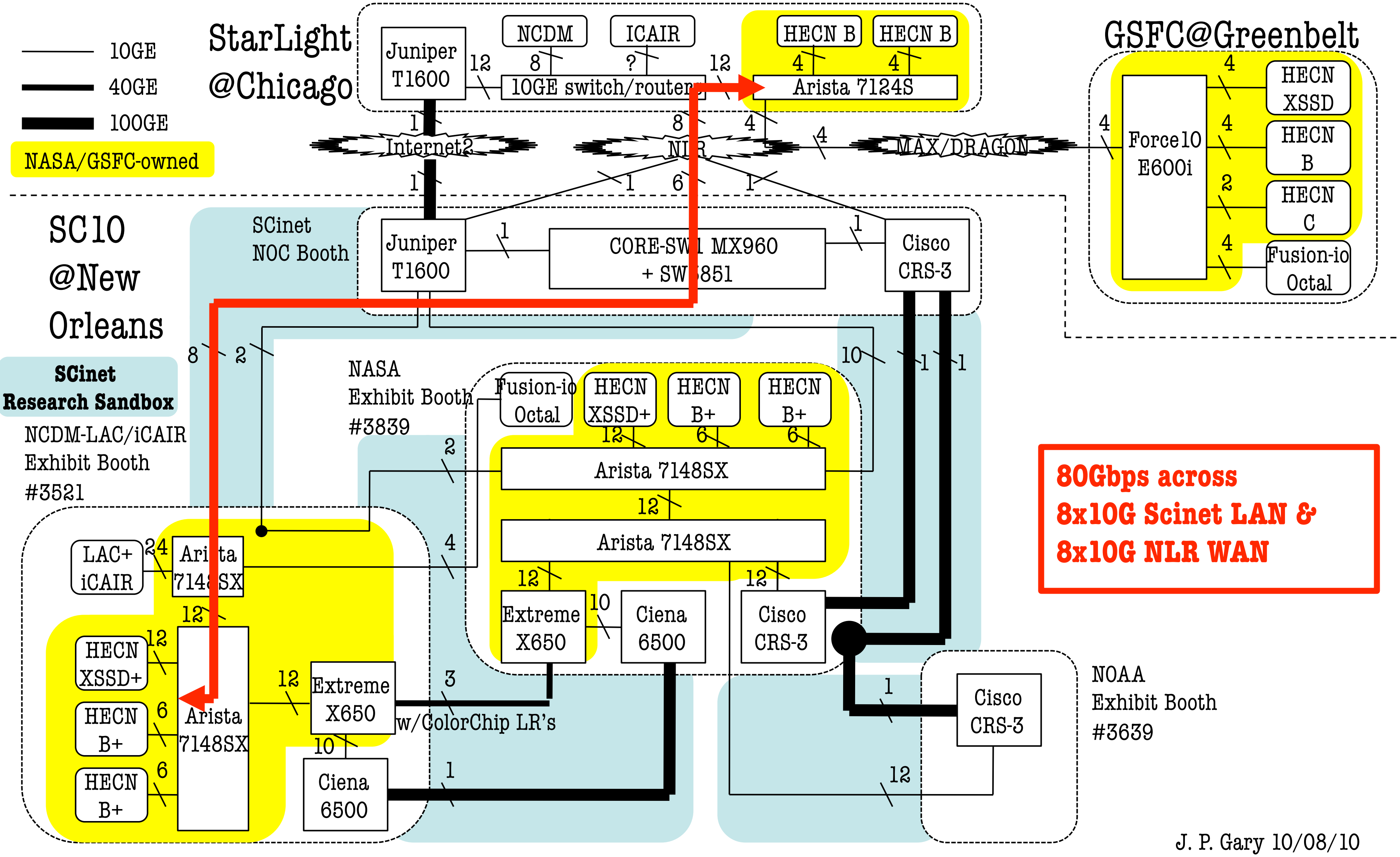


80Gbps across
8x10G Scinet LAN &
8x10G NLR WAN

J. P. Gary 10/08/10

Using 100G Network Technology in Support of Petascale Science

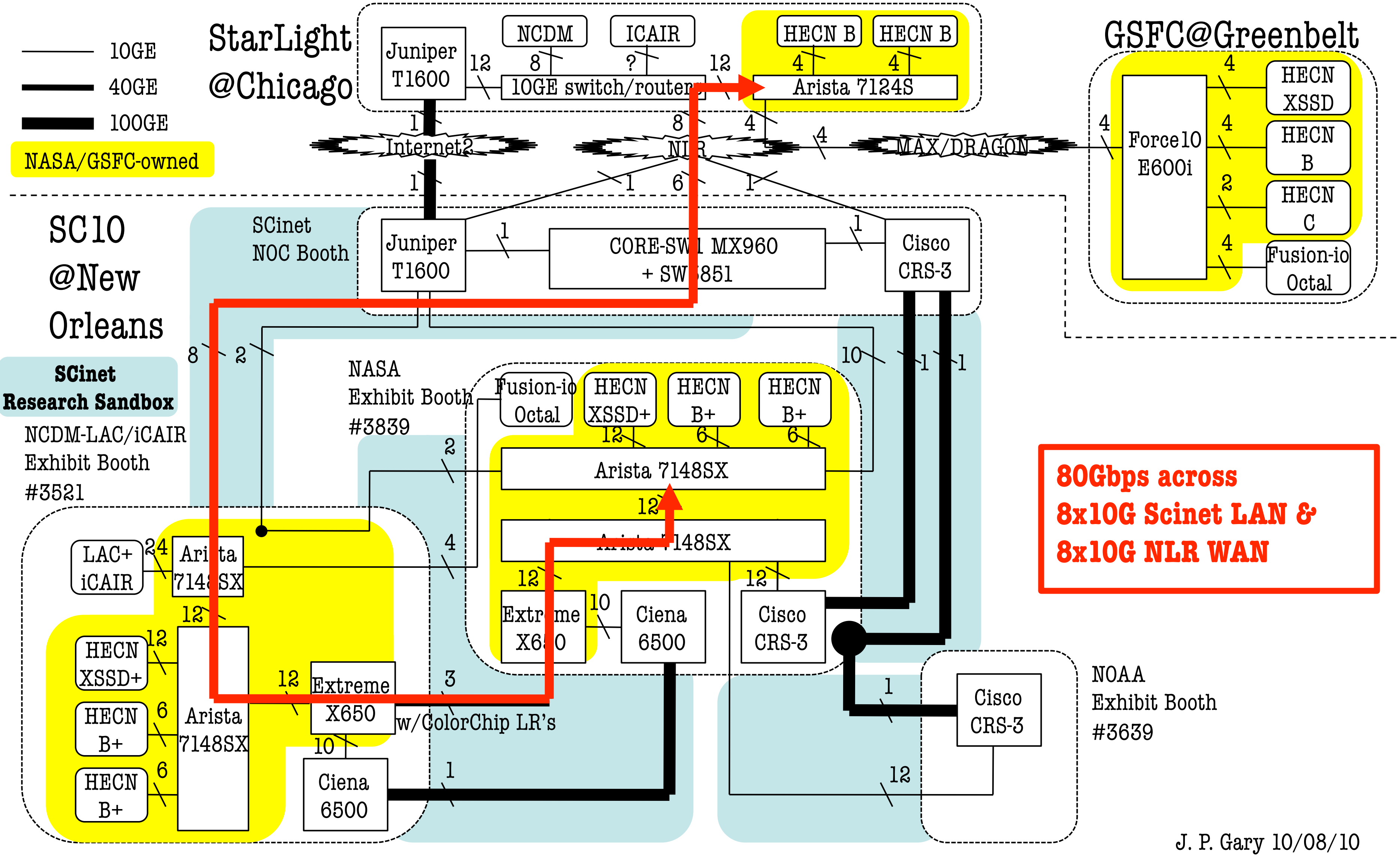
A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between Chicago and SC10



**80Gbps across
8x10G Scinet LAN &
8x10G NLR WAN**

Using 100G Network Technology in Support of Petascale Science

A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between Chicago and SC10

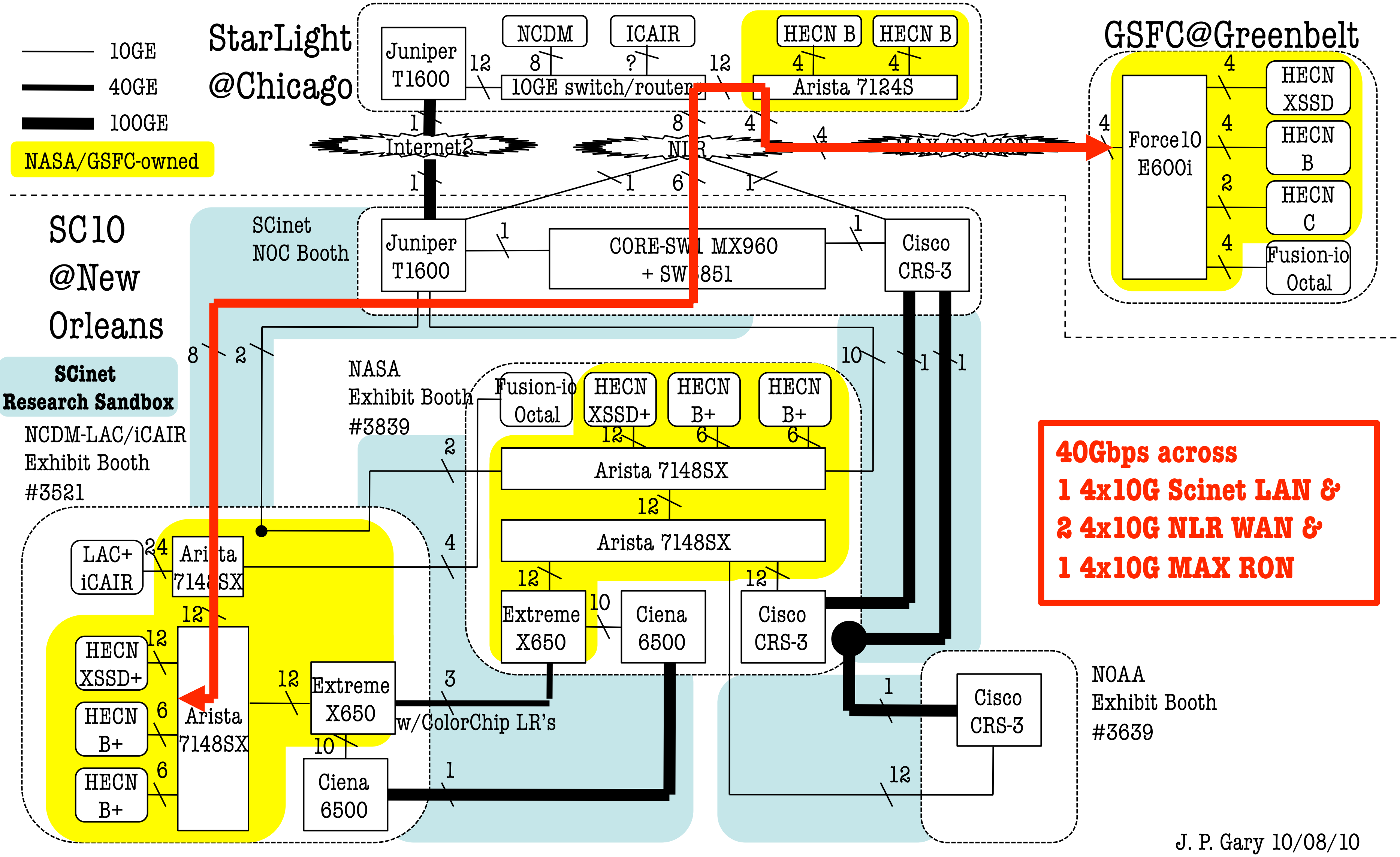


80Gbps across
8x10G Scinet LAN &
8x10G NLR WAN

J. P. Gary 10/08/10

Using 100G Network Technology in Support of Petascale Science

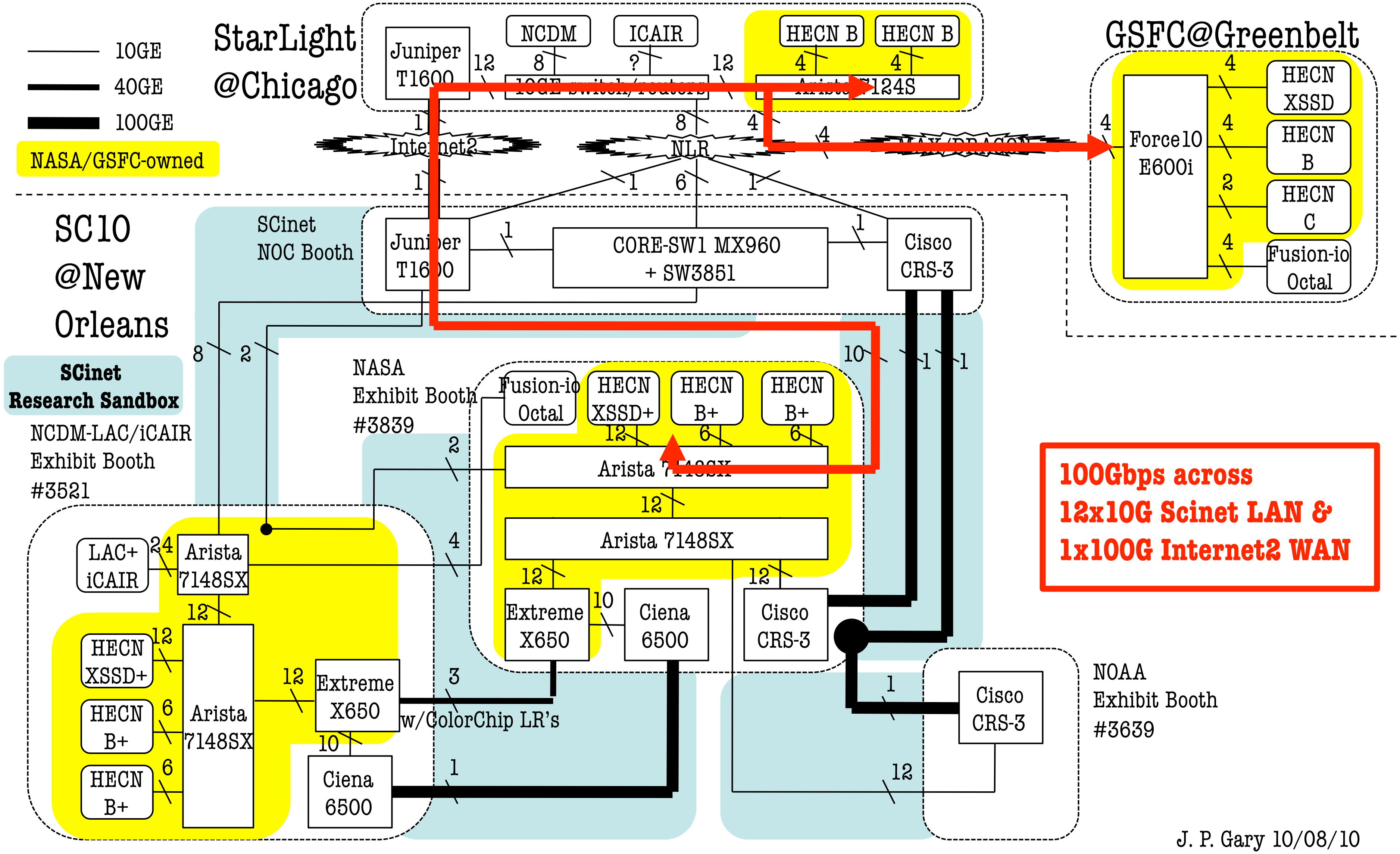
A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between Chicago and SC10



40Gbps across
1 4x10G Scinet LAN &
2 4x10G NLR WAN &
1 4x10G MAX RON

Using 100G Network Technology in Support of Petascale Science

A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between Chicago and SC10



100Gbps across 12x10G Scinet LAN & 1x100G Internet2 WAN

J. P. Gary 10/08/10

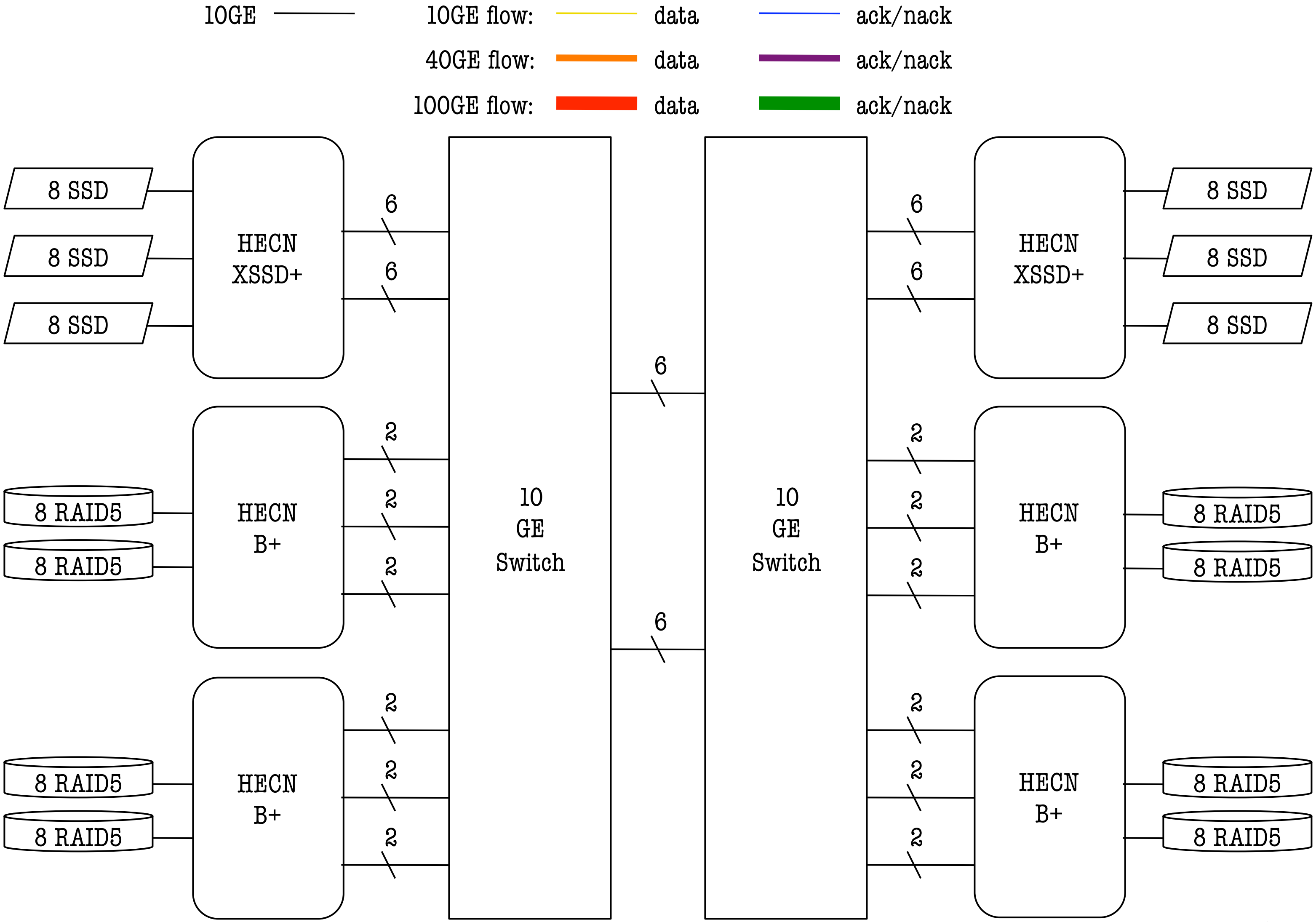
Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC10

Nuttcp >100 Gbps Uni-Directional Memory-to-Memory Flow Paths

- Receive limitations in HECN’s Xeon workstations cause use of two Core i7 workstations to support >100 Gbps uni-directional memory-to-memory data transmissions from one Xeon transmitter

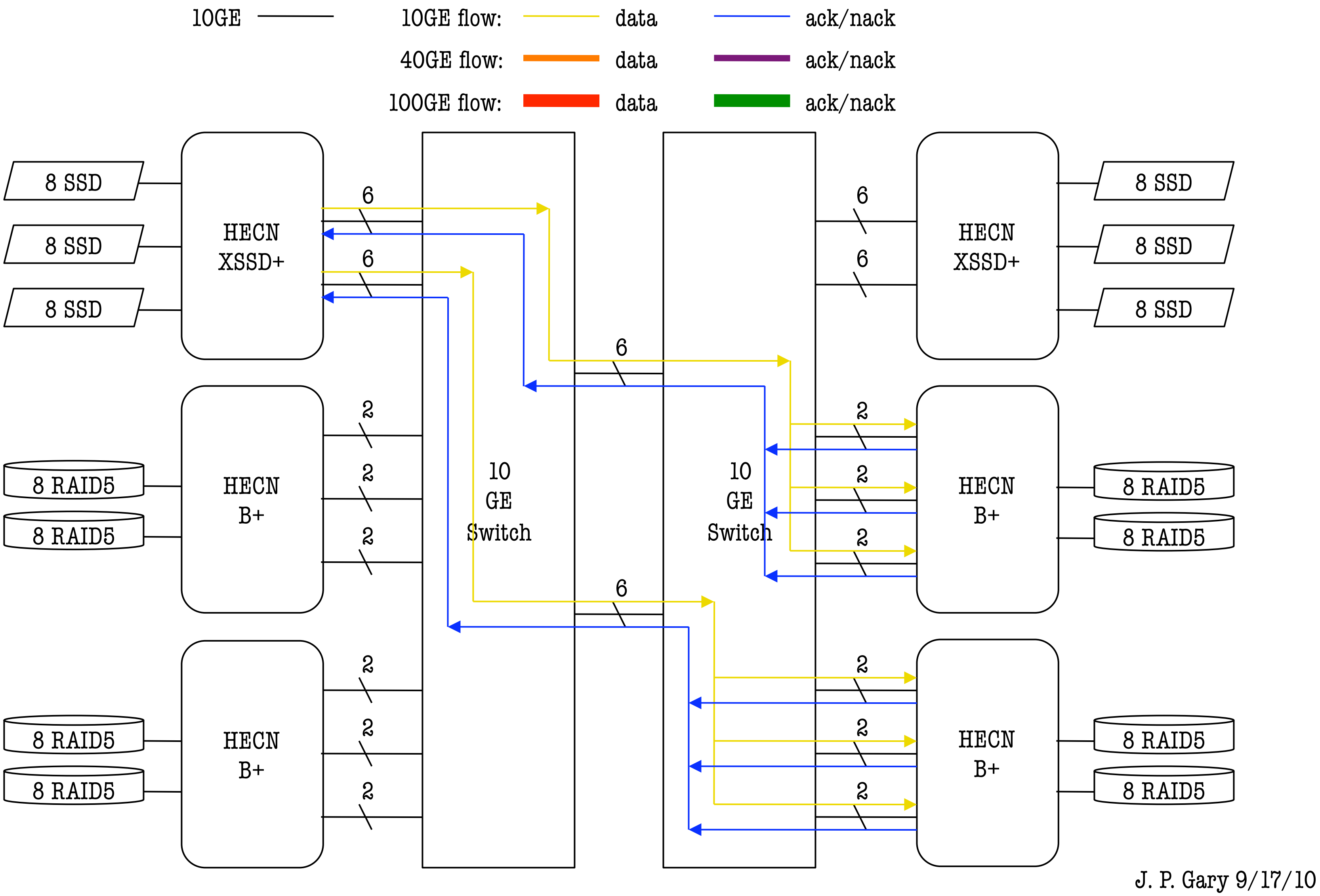
- Bi-directionally filling an intervening network with 100G data flows requires only two sets each consisting of one Xeon transmitter and two Core i7 receivers

Nuttcp >100 Gbps Uni-Directional Memory-to-Memory Flow Paths (1 of 9)

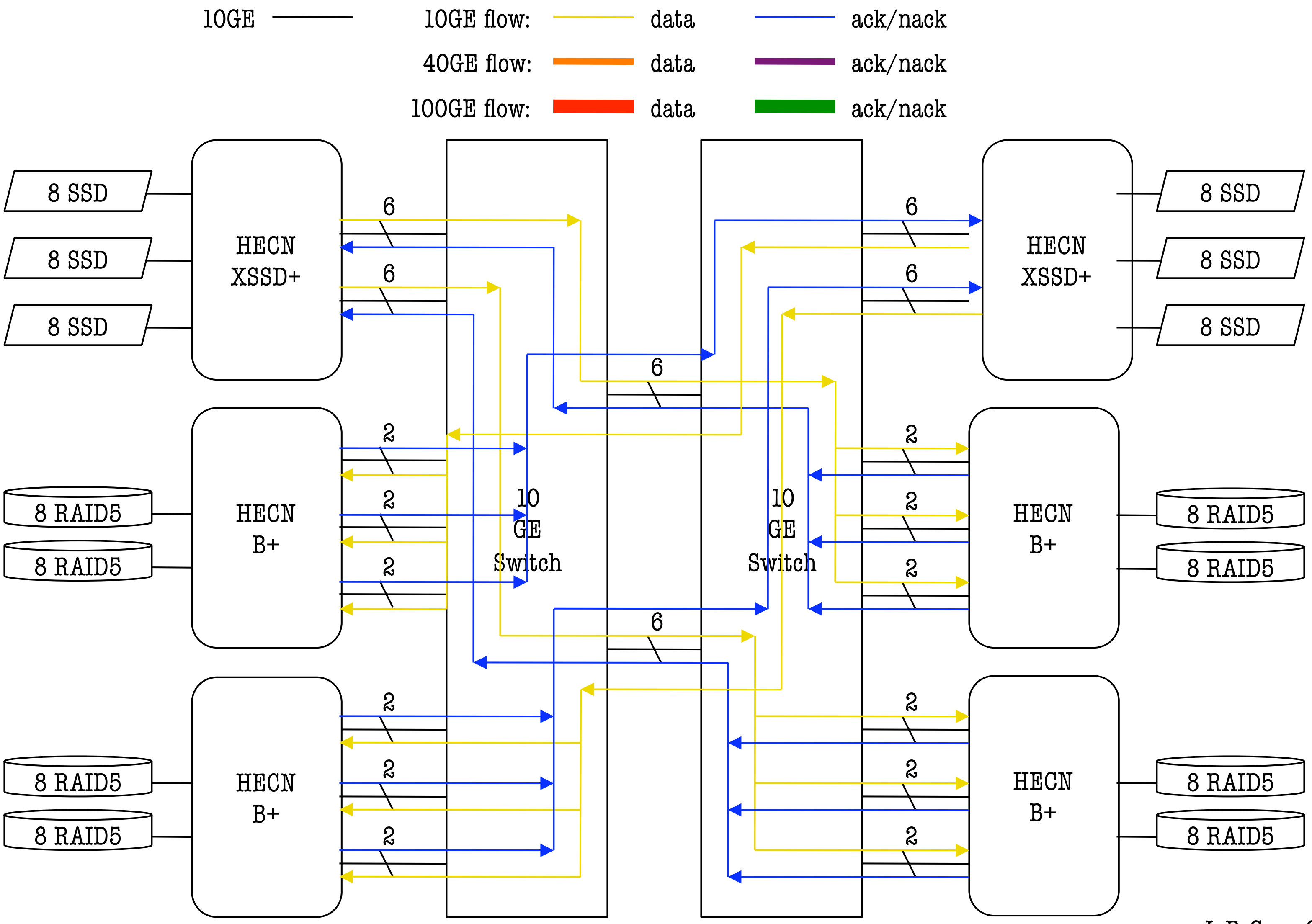


J. P. Gary 9/17/10

Nuttcp >100 Gbps Uni-Directional Memory-to-Memory Flow Paths (2 of 9)

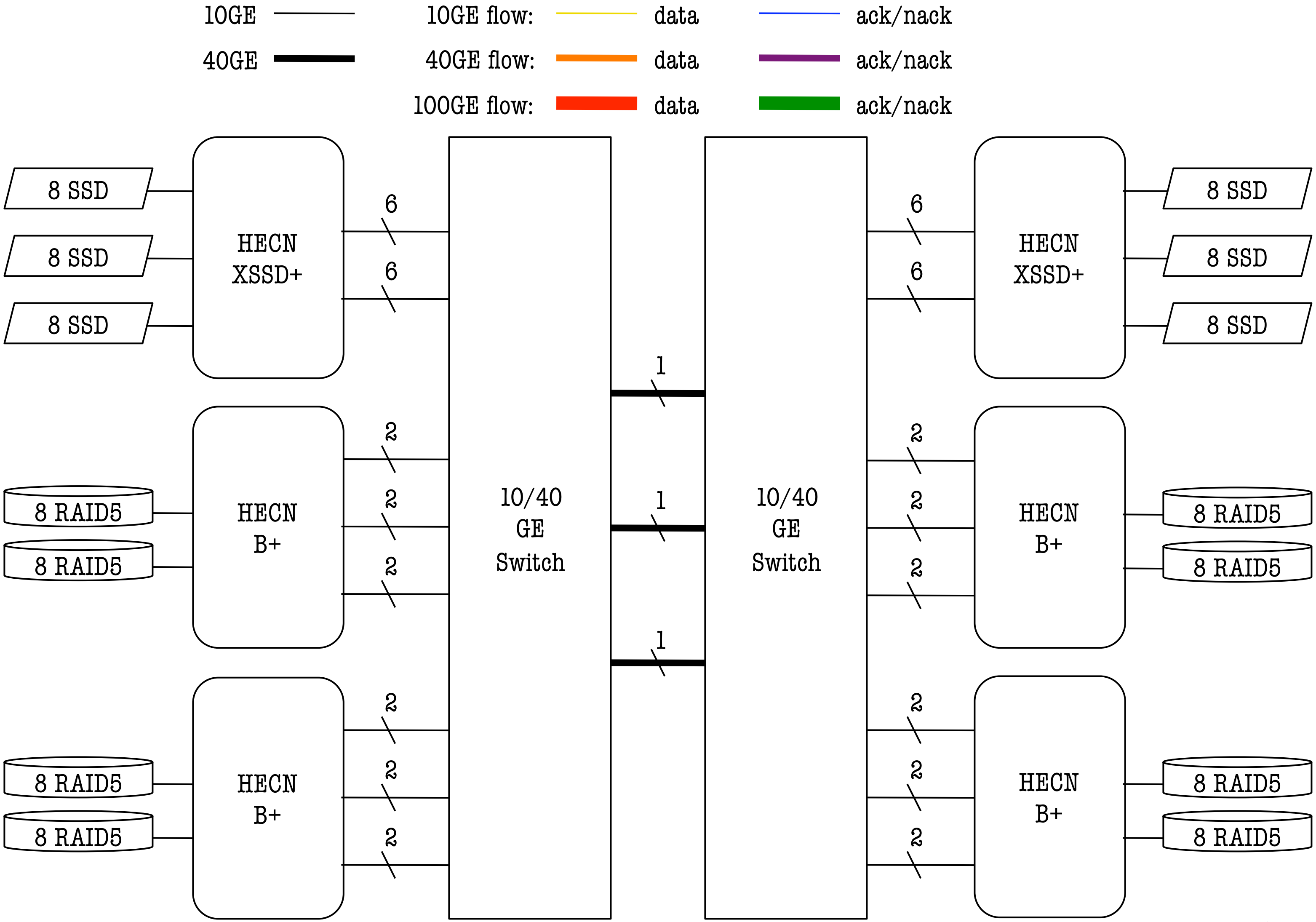


Nuttcp >100 Gbps Uni-Directional Memory-to-Memory Flow Paths (3 of 9)



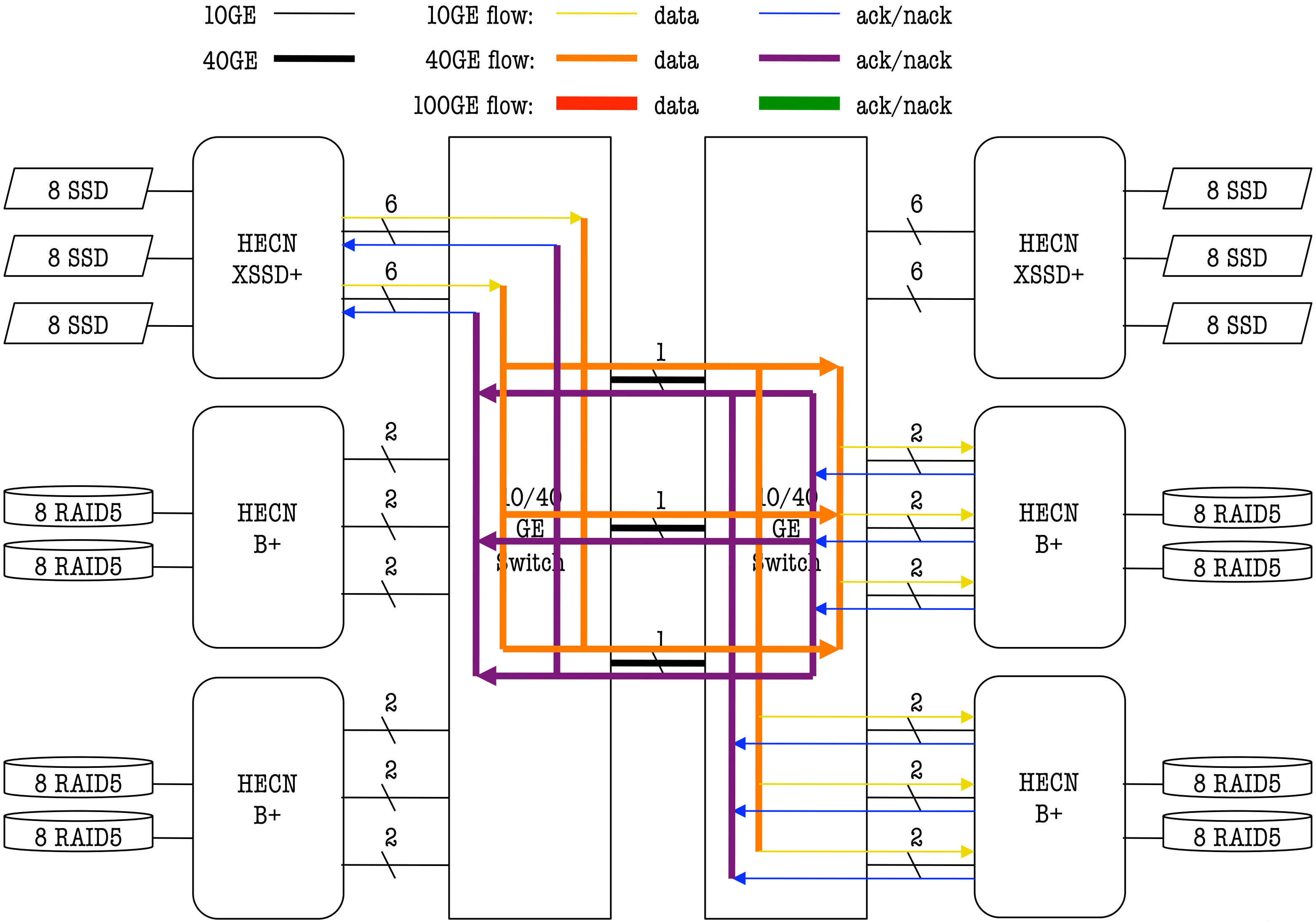
J. P. Gary 9/17/10

Nuttcp >100 Gbps Uni-Directional Memory-to-Memory Flow Paths (4 of 9)



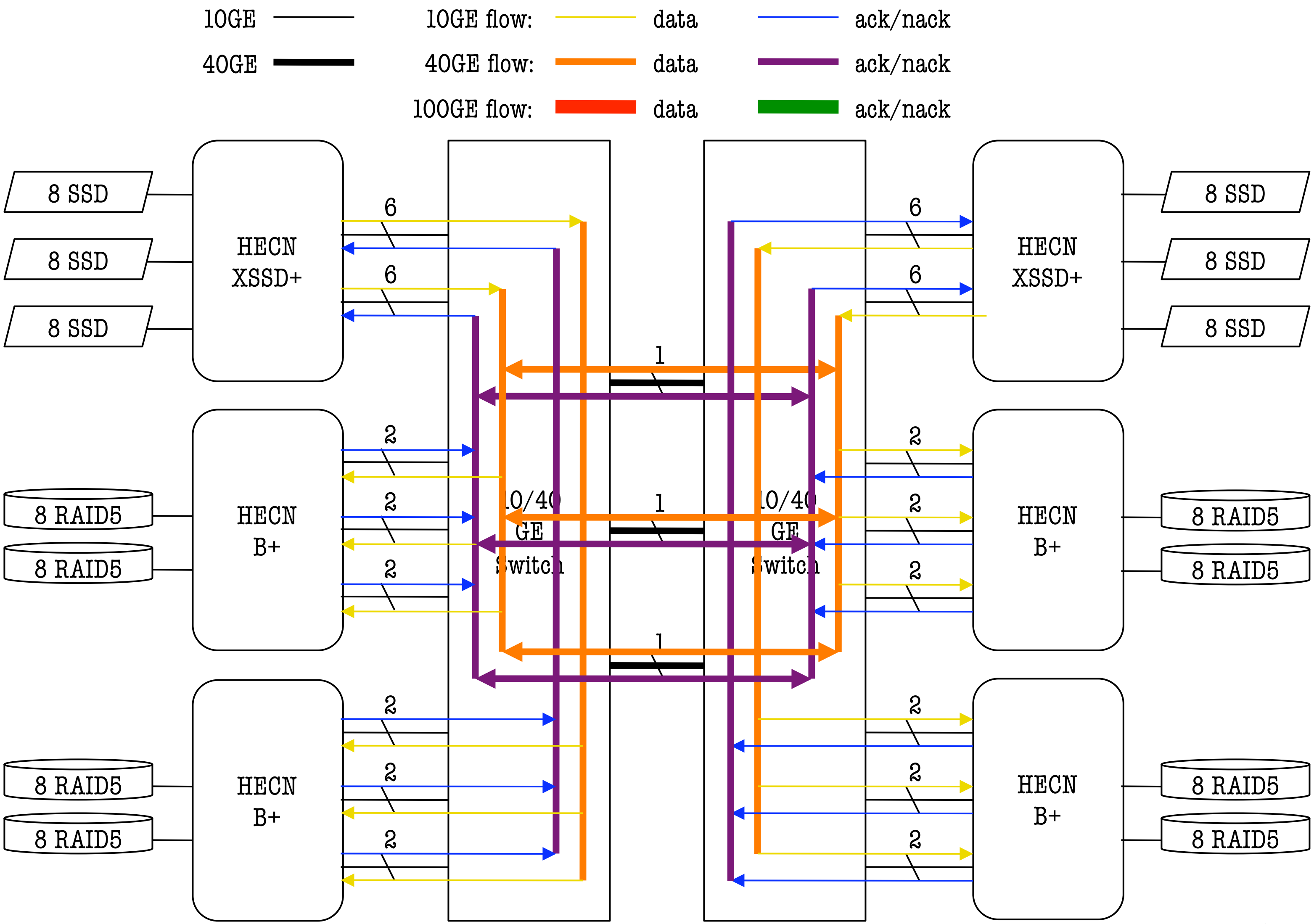
J. P. Gary 9/17/10

Nuttcp >100 Gbps Uni-Directional Memory-to-Memory Flow Paths (5 of 9)



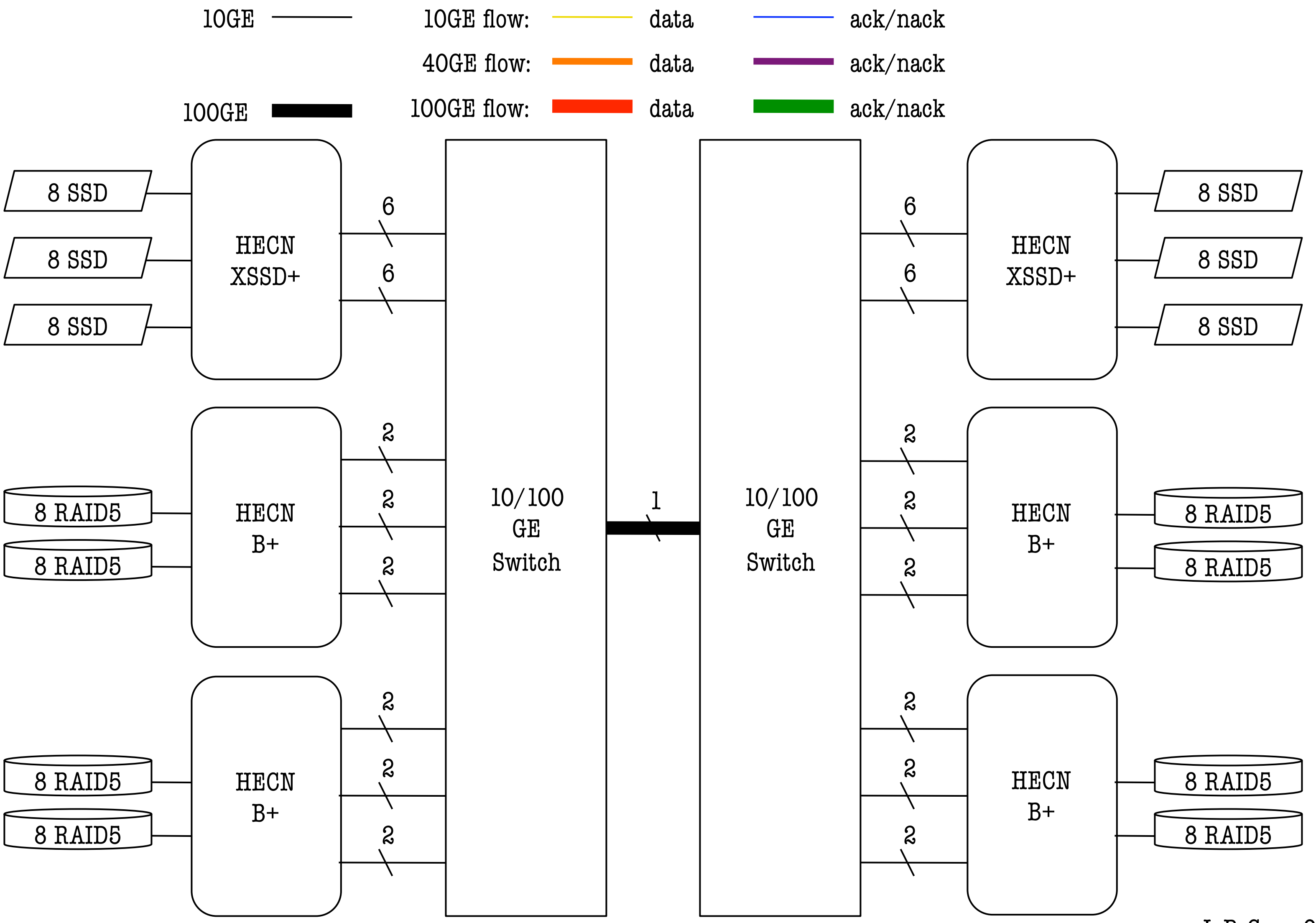
J. P. Gary 9/17/10

Nuttcp >100 Gbps Uni-Directional Memory-to-Memory Flow Paths (6 of 9)



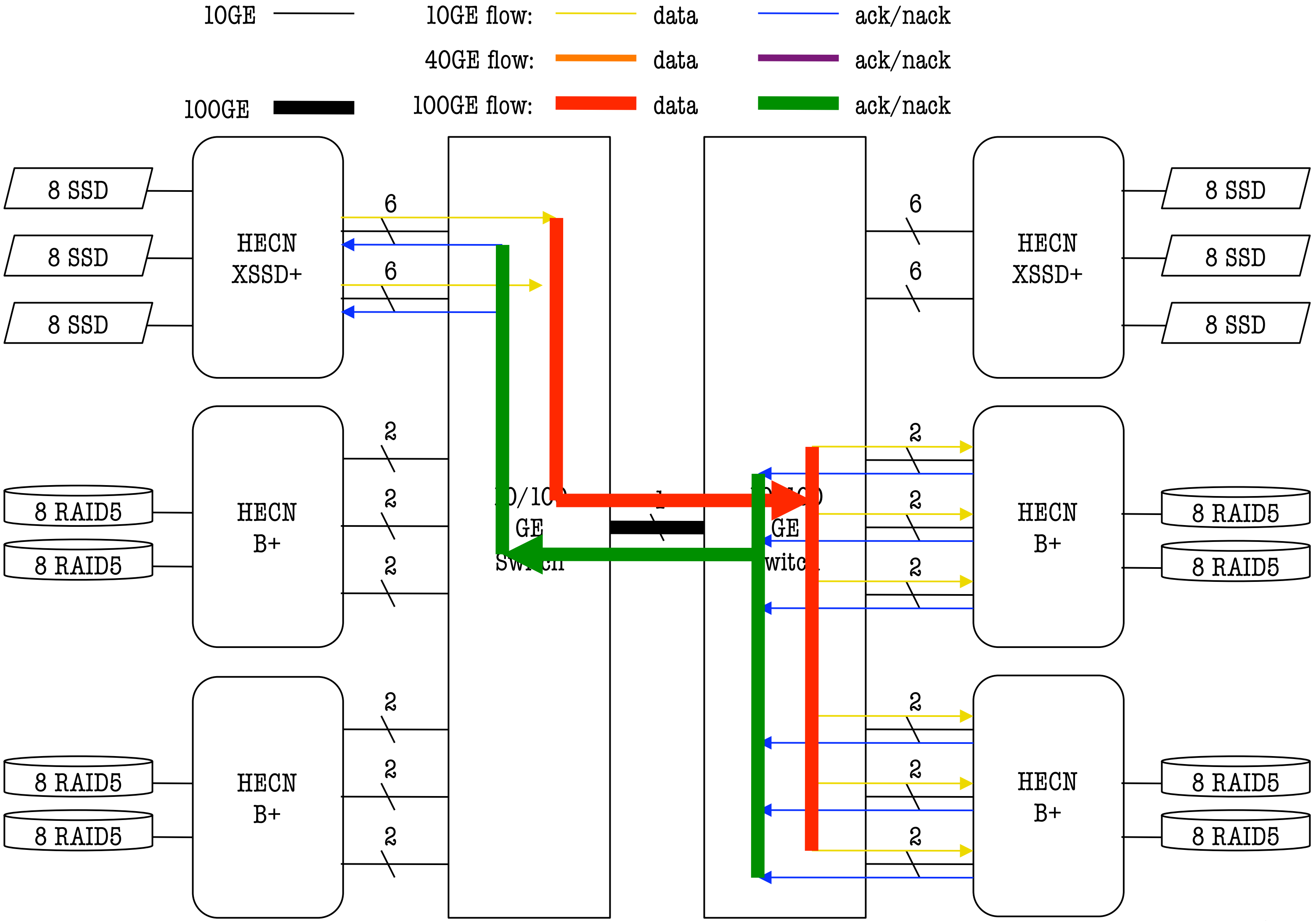
J. P. Gary 9/17/10

Nuttcp >100 Gbps Uni-Directional Memory-to-Memory Flow Paths (7 of 9)



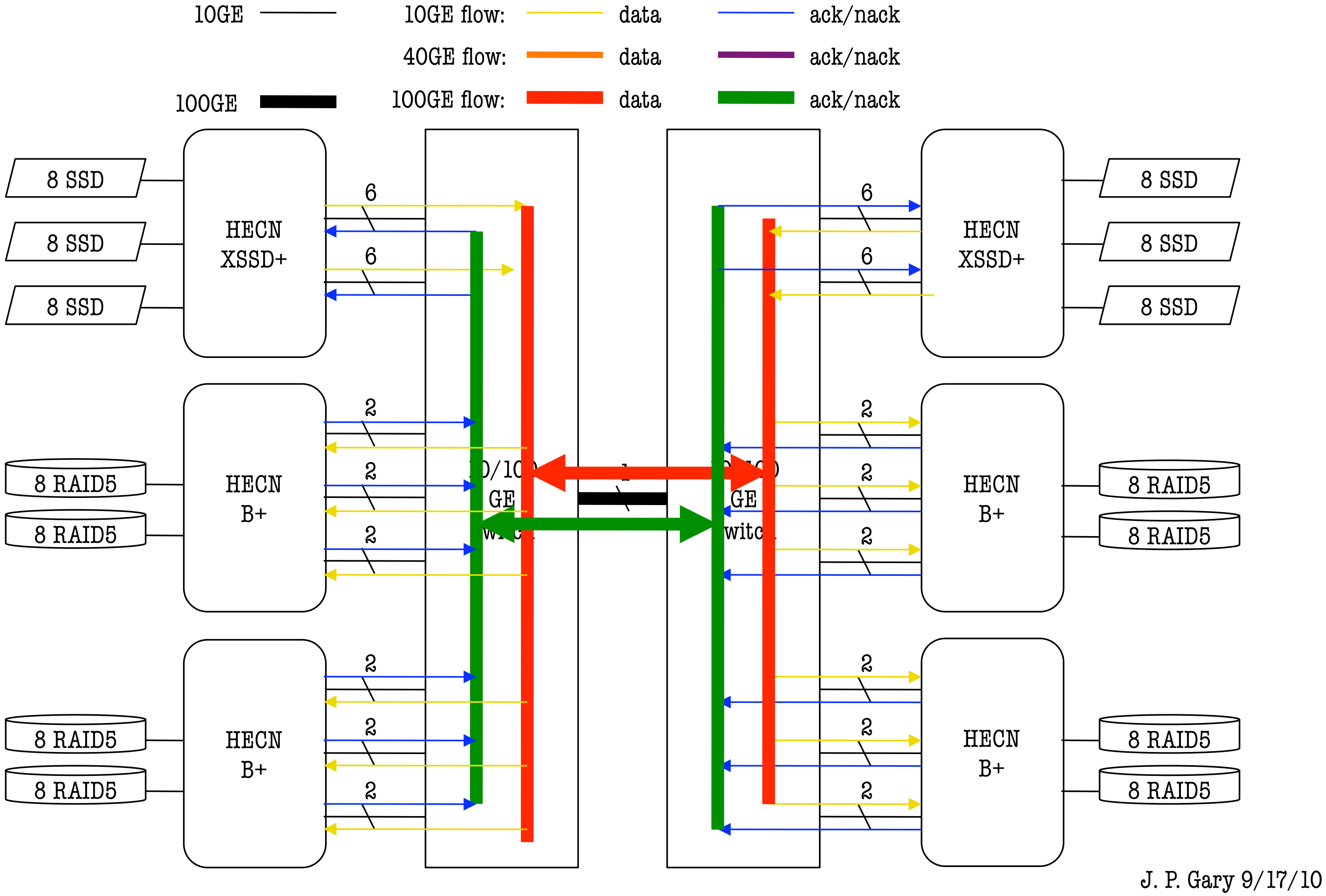
J. P. Gary 9/17/10

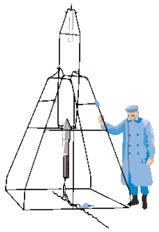
Nuttcp >100 Gbps Uni-Directional Memory-to-Memory Flow Paths (8 of 9)



J. P. Gary 9/17/10

Nuttcp >100 Gbps Uni-Directional Memory-to-Memory Flow Paths (9 of 9)





Introduction To
NASA HEC WAN File Accessing
Experiments/Demonstrations At SC10

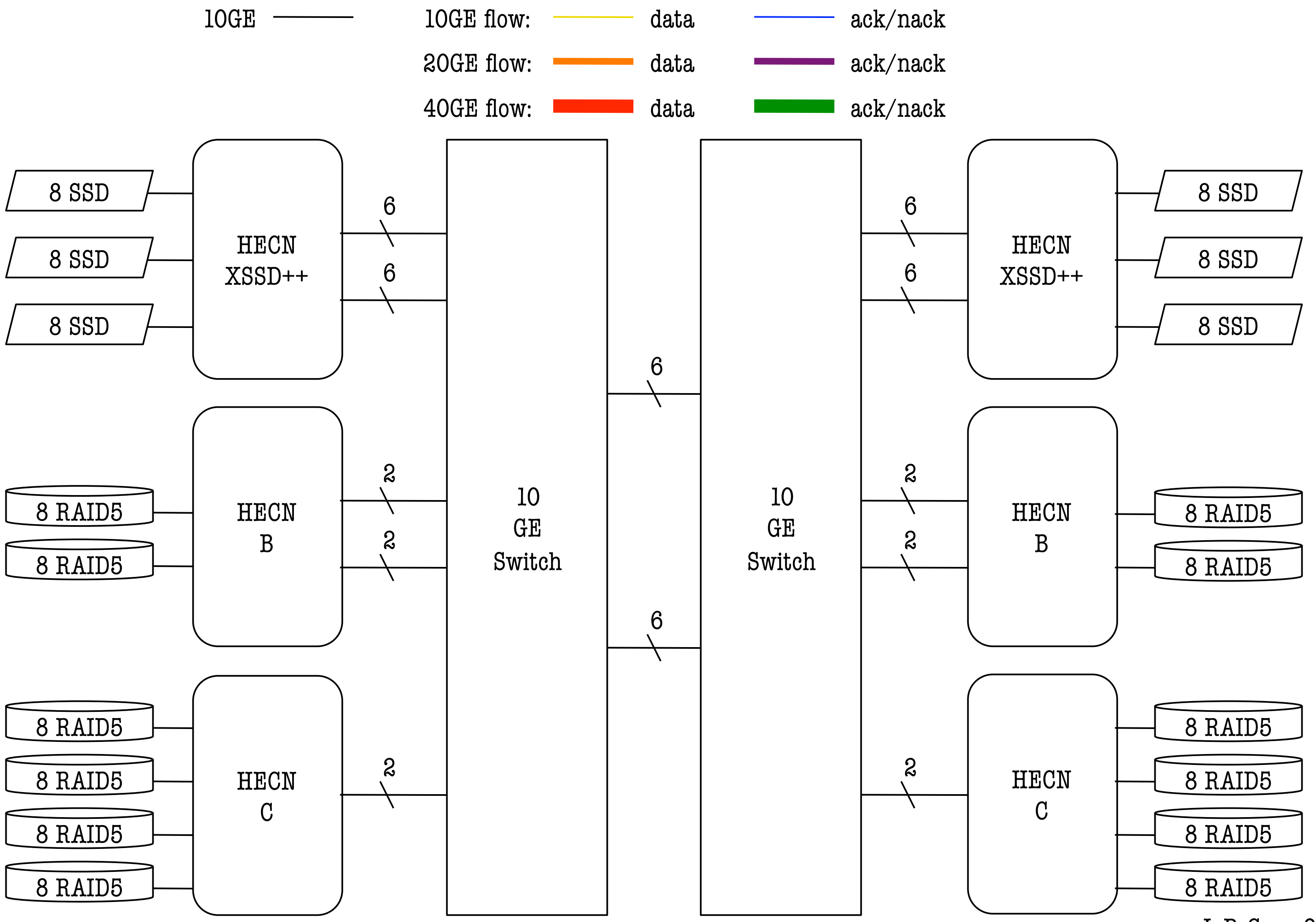
Nuttscp Near-40 Gbps Uni-Directional Disk-to-Disk Flow Paths

- Limitations in HECN’s “B” (quad Core i7) workstations (“first generation”) that achieved 10 Gbps disk-to-disk primarily were caused by too few PCIe slots (allowing only two RAID5 controllers with 16 rotating disc) and the “RAID5-cpus” in the controllers

- Limitations in HECN’s “C” (quad Core i7) workstations (also “first generation”but using 32 rotating discs) that achieved ~20 Gbps disk-to-disk primarily were caused by too little processing power in the Core i7’s

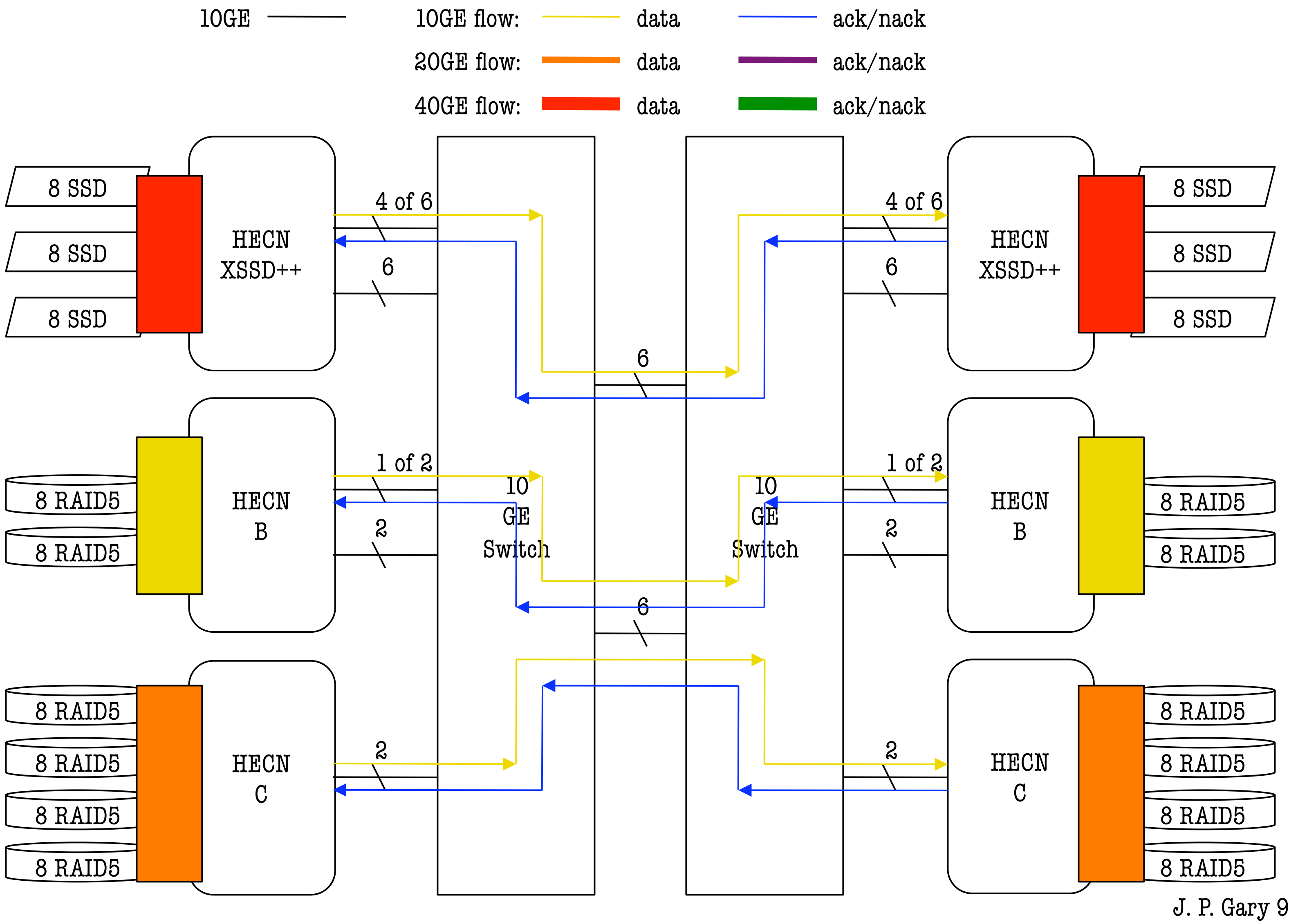
- HECN’s “XSSD” (quad Xeon) workstations (“second generation”) are targeting ~40 Gbps disk-to-disk with 24 SSDs

Nuttscp Near-40 Gbps Uni-Directional Disk-to-Disk Flow Paths (1 of 4)

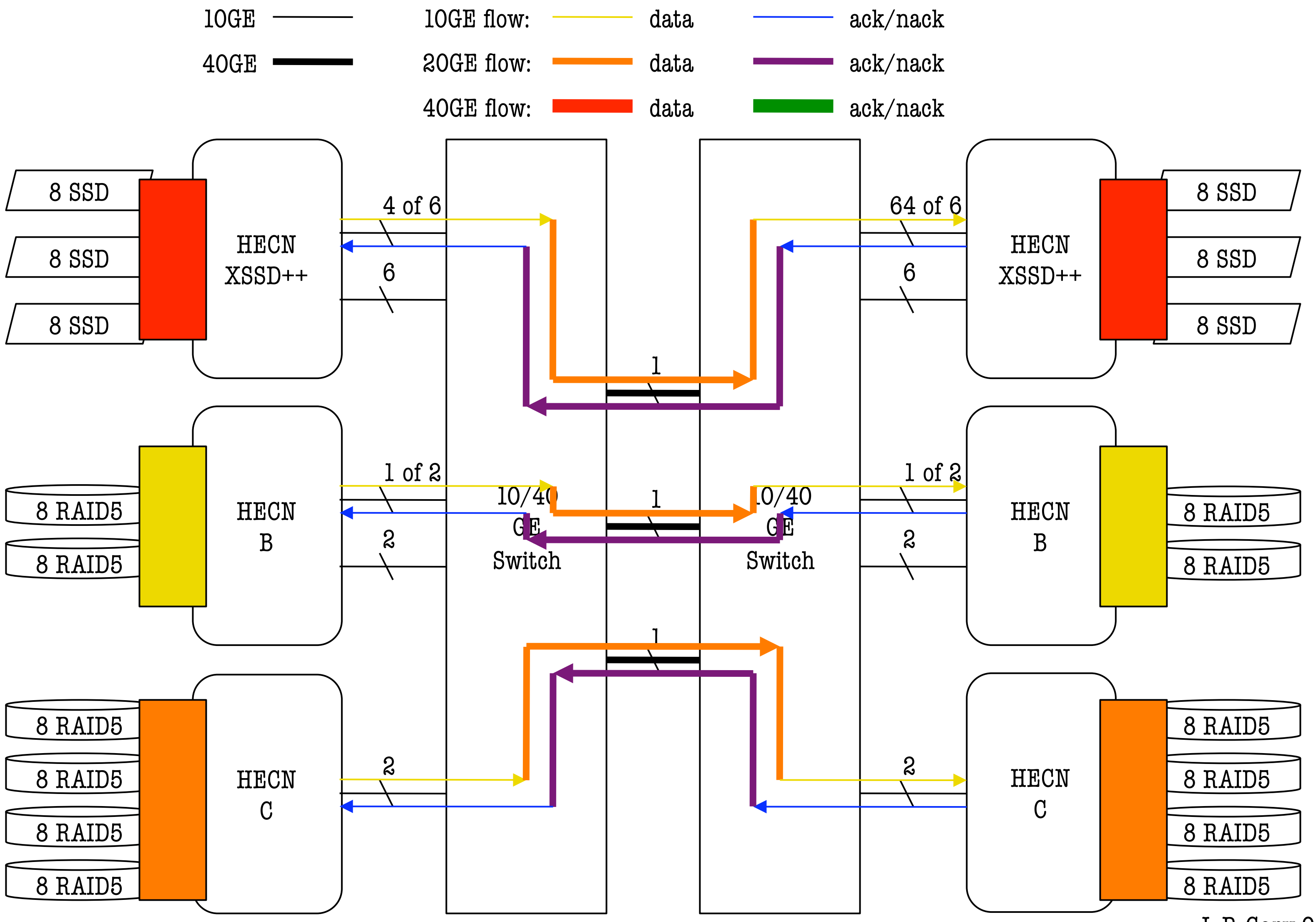


J. P. Gary 9/17/10

Nuttscp Near-40 Gbps Uni-Directional Disk-to-Disk Flow Paths (2 of 4)

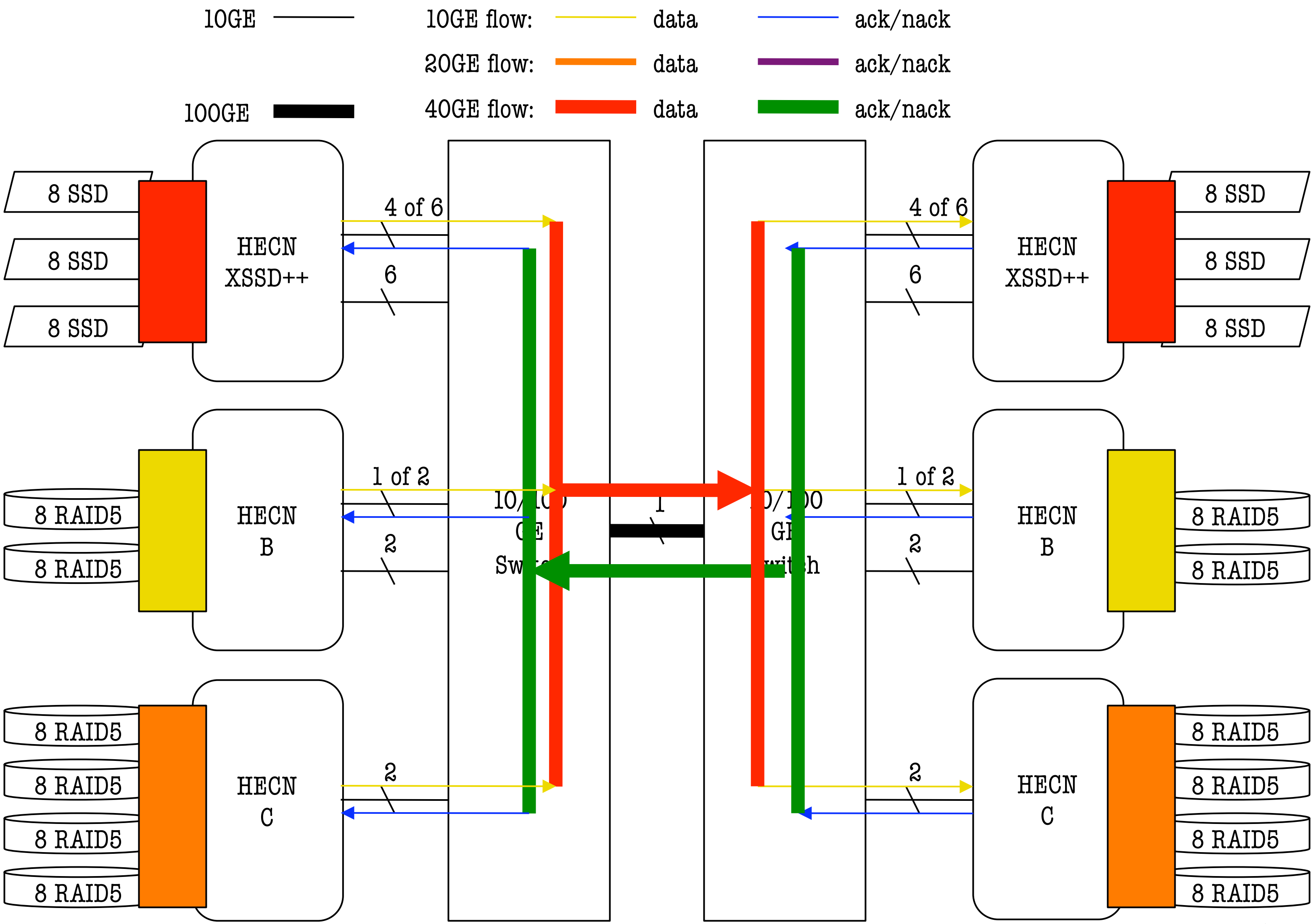


Nuttscp Near-40 Gbps Uni-Directional Disk-to-Disk Flow Paths (3 of 4)



J. P. Gary 9/17/10

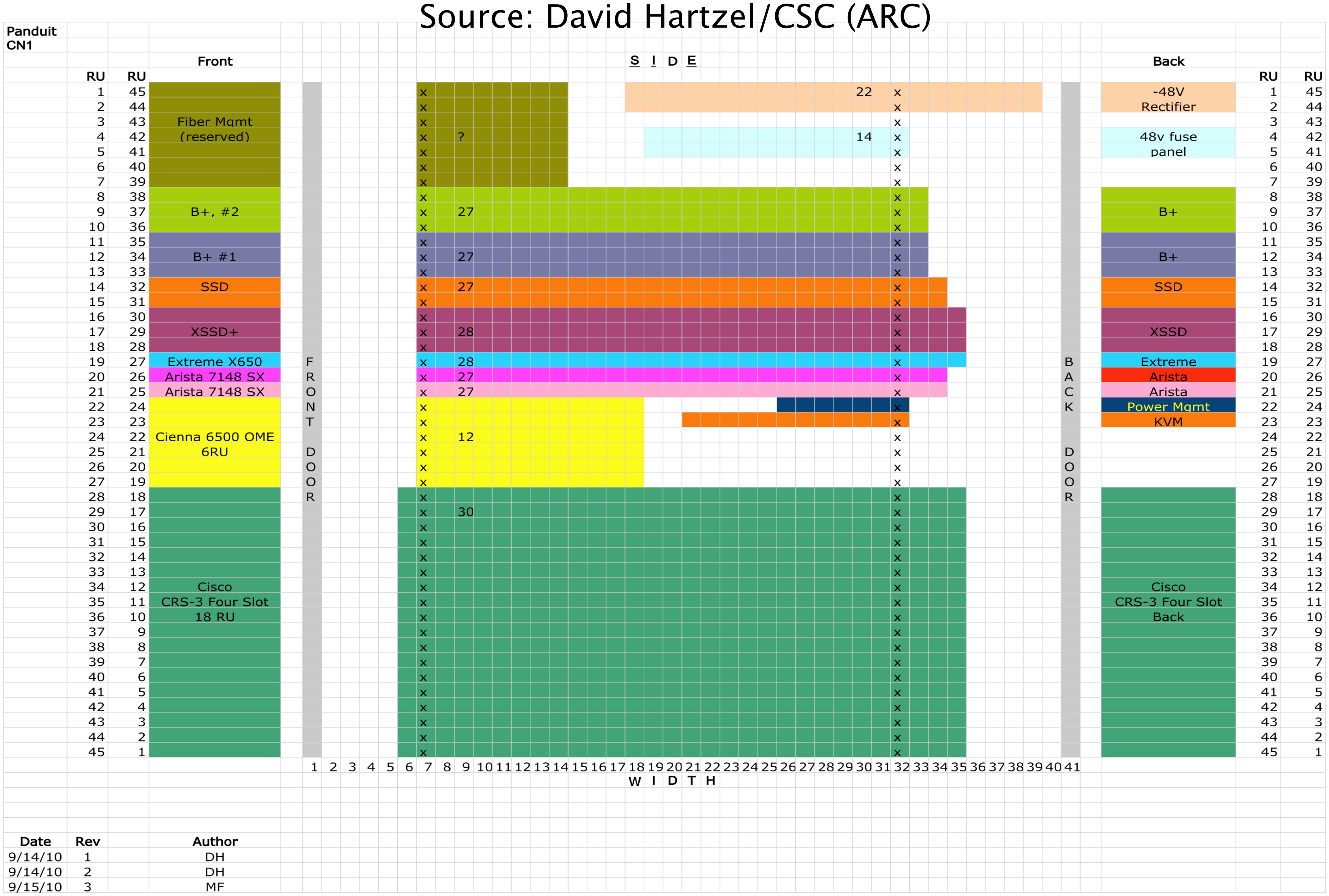
Nuttscp Near-40 Gbps Uni-Directional Disk-to-Disk Flow Paths (4 of 4)



J. P. Gary 9/17/10

Currently Proposed Layout of Equipment in Panduit Rack in NASA's Exhibit Booth

Source: David Hartzel/CSC (ARC)



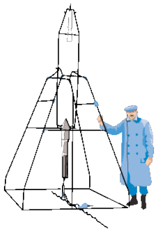
Panduit
CN1

RU	RU	Front		SIDE		Back	RU	RU
1	45	Fiber Mgmt (reserved)		x 22 x		-48V Rectifier	1	45
2	44			x x			2	44
3	43			x x			3	43
4	42			x ? 14 x		48v fuse panel	4	42
5	41			x x			5	41
6	40			x x			6	40
7	39			x x			7	39
8	38	B+, #2		x x		B+	8	38
9	37			x 27 x			9	37
10	36			x x			10	36
11	35	B+ #1		x x		B+	11	35
12	34			x 27 x			12	34
13	33			x x			13	33
14	32	SSD		x 27 x		SSD	14	32
15	31			x x			15	31
16	30	XSSD+		x x		XSSD	16	30
17	29			x 28 x			17	29
18	28			x x			18	28
19	27	Extreme X650	F	x 28 x	B	Extreme	19	27
20	26	Arista 7148 SX	R	x 27 x	A	Arista	20	26
21	25	Arista 7148 SX	O	x 27 x	C	Arista	21	25
22	24	Cienna 6500 OME 6RU	N	x x	K	Power Mgmt	22	24
23	23		T	x x		KVM	23	23
24	22			x 12 x			24	22
25	21		D	x x	D		25	21
26	20		O	x x	O		26	20
27	19		O	x x	O		27	19
28	18	Cisco CRS-3 Four Slot 18 RU	R	x x	R	Cisco CRS-3 Four Slot Back	28	18
29	17			x 30 x			29	17
30	16			x x			30	16
31	15			x x			31	15
32	14			x x			32	14
33	13			x x			33	13
34	12			x x			34	12
35	11			x x			35	11
36	10			x x			36	10
37	9			x x			37	9
38	8			x x			38	8
39	7			x x			39	7
40	6			x x			40	6
41	5			x x			41	5
42	4			x x			42	4
43	3			x x			43	3
44	2			x x			44	2
45	1			x x			45	1

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41

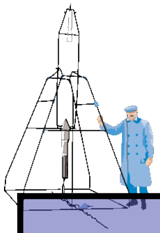
WIDTH

Date	Rev	Author
9/14/10	1	DH
9/14/10	2	DH
9/15/10	3	MF



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC10

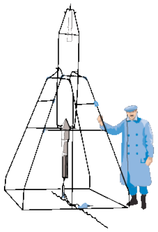
Backup Slides



Test Results Pre-3Nov09 (pre-SC09)

Source: Hoot Thompson/PTP (GSFC/NCCS)

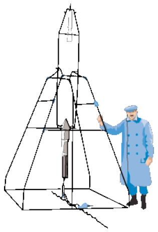
		rtt		
Tool	Type	0 msec	100 msec	Comments
nuttcp	Memory ↔ Memory	982 MB/s	920 MB/s	With large rtt, performance builds to peak number
perftest	Memory ↔ Memory	937 MB/s	N/A	rdma_bw test over 10GE NetEffect NICS
rdmacp	Disk ↔ Disk	824 MB/s	~800 MB/s	
bbftp	Disk ↔ Disk	814 MB/s (put) 840 MB/s (get)	33 MB/s (put) 33 MB/s (get)	
iRODS	Disk ↔ Disk	378 MB/s (iput) 379 MB/s (iget)	112 MB/s (iput) 43 MB/s (iget)	
xdd copy	Disk ↔ Disk	981 MB/s (src) 620 MB/s (dest)	493 MB/s (src) 372 MB/s (dest)	Added security related information
dsync	Disk ↔ Disk	N/A	N/A	rdma rsync – just now available
nuttscp	Disk ↔ Disk	577 MB/s	577 MB/s	Default settings
nfs	Disk ↔ Disk	686 MB/s (wrt) 444 MB/s (read)	Not Useful	
nfsrdma	Disk ↔ Disk	319 MB/s (wrt) 326 MB/s (read)	Not Useful	Could not achieve advertised results



Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Nuttcp (pronounced as new-t-t-c-p or nut-t-c-p)

- Primary author Bill Fink (william.e.fink@nasa.gov), with Rob Scott (rob@hpcmo.hpc.mil).
- Great follow-on to original ttcp network throughput performance measurement and troubleshooting tool. Highly recommended. One of the best!
- Over 60 examples of use included in Phil Dykstra's noteworthy tutorial for High Performance Data Transfer (at SC0x's).
- Advanced capabilities/features/options still being added to enable more sophisticated use, while retaining ease-of-use defaults.
- At http://www.nuttcp.net & included in perfSONAR's liveCD.



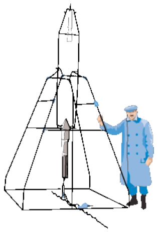
Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Phase 1 Network-Test Workstation Functional Objectives with Performance Targets (1 of 2)

- “B” (Baseline) systems:
  - Primarily for network throughput evaluations via nuttcp memory-to-memory testing at up to 40-Gbps unidirectional, 40-Gbps bidirectional (80-Gbps “total”)
  - Secondarily for WAN file copying application throughput evaluations in disk-to-disk testing at up to 10-Gbps unidirectional
- “C” systems:
  - Primarily for WAN file copying application throughput evaluations in disk-to-disk testing at up to 20-Gbps unidirectional
- “A” systems:
  - Primarily for WAN delay emulation at up to 40-Gbps unidirectional, 40-Gbps bidirectional (80-Gbps “total”)
  - Also as firewall at up to 20-Gbps unidirectional, 20-Gbps bidirectional (40-Gbps “total”)

Phase 1 Network-Test Workstation Functional Objectives with Performance Targets (2 of 2)

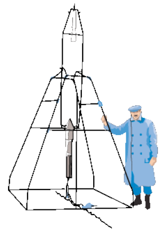
- “A+” systems:
  - Primarily for network throughput evaluations via nuttcp memory-to-memory testing at up to 70-Gbps unidirectional, 40-Gbps bidirectional (80-Gbps “total”)
    - Actual performance: On 12Jun09 using eight streams between two A+ systems connected via eight 10GE’s, measured an aggregate performance of 69.2907 Gbps unidirectional, and bidirectional 38.6955 Gbps transmit & 38.5842 Gbps receive (77.2797 Gbps total aggregate)
- “A-” systems:
  - Primarily for network throughput evaluations via nuttcp memory-to-memory testing at up to 20-Gbps unidirectional, 20-Gbps bidirectional (40-Gbps “total”)



Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Phase 1.1 Network-Test Workstation Functional Objectives with Performance Targets

- “X++” systems:
  - Primarily for network throughput evaluations via nuttcp memory-to-memory testing at up to 100-Gbps unidirectional, 50-Gbps bidirectional (100-Gbps “total”)
- Actual performance “in-progress”
  - On 6Aug09 measured an aggregate performance of 100.4637 Gbps in transmits; but currently only up to 56.4703 Gbps in receives
    - Test configuration has each of the two quad-core Xeon processors of one X++ system connected via six 10GE’s to one of two quad-core i7-based A+ systems
    - Twelve streams are generated – one for each of the twelve 10GE connections handled by the one X++ system

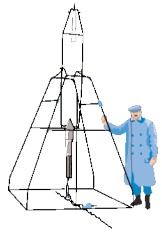


Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Approximate Costs (With components acquired via SEWP IV in lot-sizes of 3 - 15, and self assembly) of Phase 1 & 1.1 Network-Test Workstations

“B” System:	~$6.8K
“C” System:	~$9.0K
“A” System:	~$4.6K
“A+” System:	~$6.5K
“A-” System:	~$3.6K
“X++” System:	~$11.1K

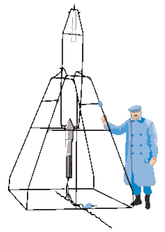
- For more detail, contact Paul.Lang@nasa.gov



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Nuttscp Sample Test Results Between Two "B-Systems" (1 of 4) [Source: Bill Fink/GSFC]

- Two simultaneous 64-GB file copies (each file-copy streamed between one RAID5 disk controller hosted on each B-system in a LAN testbed)
  - File copy 1: 5092.5196-Mbps 43% TX 77% RX 0 retrans 0.10ms RTT
  - File copy 2: 5045.3832-Mbps 33% TX 77% RX 0 retrans 0.10ms RTT
- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system in a LAN testbed)
  - File copy: 9824.2054-Mbps 58% TX 96% RX 0 retrans 0.10ms RTT
- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system in a 40km MAN testbed)
  - File copy: 9402.0330-Mbps 56% TX 98% RX 0 retrans 0.45ms RTT

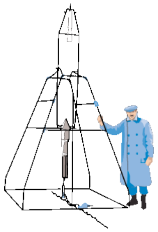


Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Nuttscp Sample Test Results Between Two “B-Systems” (2 of 4) [Source: Bill Fink/GSFC]

- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system* in a ~3000km-emulated (by netem) WAN testbed)
  - File copy: 9548.0962-Mbps 59% TX 97% RX 0 retrans 80.15ms RTT (completed in 57.58 seconds)
    *With receiver B-system over-clocked to 3.4-Ghz instead of 3.2-Ghz
  - [For comparison a 60.16 second memory-to-memory test using nuttcp:
    9661.2217-Mbps 26% TX 40% RX 0 retrans 80.14ms RTT]

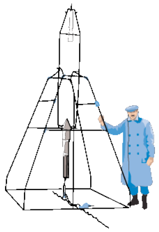
- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system** in a ~3000km-emulated (by netem) WAN testbed)
  - File copy: 8931.9535-Mbps 58% TX 97% RX 0 retrans 80.14ms RTT (completed in 61.55 seconds)
    **With receiver B-system clocked normally at 3.2-Ghz



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Nuttscp Sample Test Results Between Two "B-Systems" (3 of 4) [Source: Bill Fink/GSFC]

- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system* in a ~3000km-emulated (by netem) WAN testbed)
  - File copy: 5055.1438-Mbps 31% TX 59% RX 8 retrans 80.15ms RTT (completed in 108.75 seconds)
    *With receiver B-system over-clocked to 3.4-Ghz instead of 3.2-Ghz
  - [For comparison a 30.29 second memory-to-memory test using nuttcp:
    5561.7408-Mbps 14% TX 28% RX 4 retrans 80.15ms RTT]
  - Retrans caused by "dropped_bad_crc32" errors at ~10^-6 packet loss rate



Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Nuttscp Sample Test Results Between Two "B-Systems" (4 of 4) [Source: Bill Fink/GSFC]

- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system* in a ~3000km real WAN testbed): GSFC→ARC
  - File copy: 7575.1083-Mbps 47% TX 89% RX 0 retrans 80.58ms RTT (completed in 72.57 seconds)
    *With receiver B-system clocked normally at 3.2-Ghz
- One 64-GB file copy (between two RAID5 disk controllers nested as RAID50 hosted on each B-system** in a ~3000km real WAN testbed): ARC→GSFC
  - File copy: 8284.2127-Mbps 60% TX 95% RX 0 retrans 80.58ms RTT (completed in 66.36 seconds)
    **With receiver B-system clocked normally at 3.2-Ghz

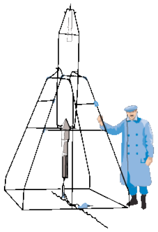
Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

Additional 10-Gbps netem-enabled-WAN Sample Test Results (nuttcp-based) Between Two "B-Systems**"* [Source: Bill Fink/GSFC]

*Also see: *Nuttscp Sample Test Results Between Two "B-Systems" (2 of 4)* & *(3 of 4)*

**With receiver B-system over-clocked to 3.4-Ghz instead of 3.2-Ghz

- With Large Receive Offload on the myri10ge driver enabled
  - 30 second test using Linux TCP autotuning:
    6053.8558-Mbps 15% TX 24% RX 0 retrans 80.14ms RTT
  - 30 second test using manually specified 100 MB TCP window:
    6796.0992-Mbps 16% TX 27% RX 0 retrans 80.15ms RTT
- With Large Receive Offload on the myri10ge driver disabled
  - 30 second test using Linux TCP autotuning:
    7029.8505-Mbps 19% TX 29% RX 0 retrans 80.15ms RTT
  - 30 second test using manually specified 100 MB TCP window:
    9442.1071-Mbps 27% TX 39% RX 0 retrans 80.15ms RTT

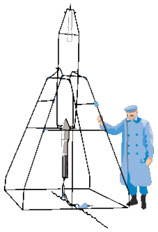


Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

<u>Sample 4x10-GigE Bonding Test Results (nuttcp-based) Between Two “B-Systems” in Back-to-Back Direct Connection</u> [Source: Bill Fink/GSFC]

- Kernal L2 load-balanced round-robin bonded interface (aka Link Aggregation)
  - 10 second test:
    <u>31615.4616-Mbps</u>   99% TX   95% RX   31 retrans   0.05ms RTT
- Nuttcp “application bonding” using 4 streams (each across its own 10-GigE path)
  - 10 second test:
    <u>39564.4536-Mbps</u>   81% TX   94% RX   0 retrans   0.11ms RTT

In both cases the use of the “correct” CPU made a significant difference in the achieved network performance. Unfortunately the “correct” CPU did not seem to be deterministic.



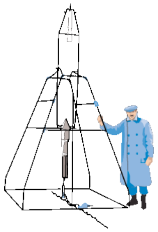
Introduction To NASA HEC WAN File Accessing Experiments/Demonstrations At SC09

More 4x10-GigE Bonding Test Results (nuttcp-based) Between Two “B-Systems”* [Source: Bill Fink/GSFC]

- Nuttcp “application bonding” using 4 streams (each across its own 10-GigE path)
  - 10 second test:
    39134.0831-Mbps 99% TX 91% RX 1 = 0+0+1+0 retrans 0.11ms RTT
  - 10 second test:
    39151.9019-Mbps 91% TX 92% RX 1 = 0+0+1+0 retrans 0.11ms RTT
  - 10 second test:
    39318.0384-Mbps 80% TX 90% RX 1 = 0+0+1+0 retrans 0.10ms RTT
  - 10 second test:
    39406.0384-Mbps 79% TX 92% RX 1 = 0+0+1+0 retrans 0.10ms RTT

*Obtained while testing nuttcp-7.1.1’s new features for:

- Improved multilink aggregation specification options (e.g., stride & dotted quad)
- Providing summary TCP retrans info for multi-stream TCP (with per-stream info for Linux)
- Allowing local name resolution to occur for third party nuttcp tests if the remote third party host can’t resolve the specified test hostname

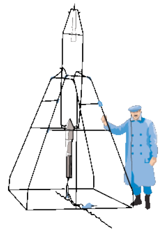


Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

Precursor Tests of “C-Systems” (to show the individual components have the necessary muscle) [Source: Bill Fink/GSFC]

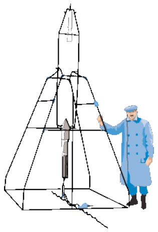
- Disk I/O speeds via dd reads (of=/dev/null) & writes (if=/dev/zero)
  - Read: 68719476736 bytes (69 GB) copied, 25.8791 s, 2.7 GB/s
  - Write: 68719476736 bytes (69 GB) copied, 26.8676 s, 2.6 GB/s

- 2x10-GigE via “nuttcp application bonding”
  - TX: 19805.8537 Mbps 34 %TX 59 %RX 0 retrans 0.11 msRTT
  - RX: 19808.7300 Mbps 39 %TX 53 %RX 0 retrans 0.11 msRTT



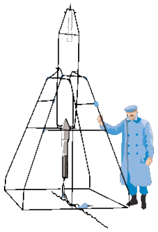
Nuttscp Sample Test Results Between Two "C-Systems" (1-of-7) [Source: Bill Fink/GSFC]

- One 64-GB file copy (between four RAID5 disk controllers nested as RAID50 hosted on each C-system in a LAN testbed)
  - Configuration settings:
    - LRO enabled
    - eth2,3 interrupts on CPU0
    - nuttcp application running on CPU1
    - 4xHPT RAID5 interrupts running on CPU2
    - md RAID50 across above
  - Get: 10273.4125 Mbps 52 %TX 99 %RX 0 retrans 0.11 msRTT
  - Put: 10311.2700 Mbps 52 %TX 99 %RX 0 retrans 0.11 msRTT
- Houston, we have a problem! We're definitely not firing on all cylinders. It's obvious what the problem is, namely that the receiver CPU is totally saturated. To go faster is going to require nuttcp using multiple cores in parallel….



Nuttscp Sample Test Results Between Two "C-Systems" (2-of-7) [Source: Bill Fink/GSFC]

- One 64-GB file copy similar to "1-of-7" but only one side's RAID50 is real
  - Configuration settings: same as in "1-of-7"
  - Get from RAID50 to /dev/null:
    17324.4416 Mbps 98 %TX 49 %RX 0 retrans 0.11 msRTT
  - Put from /dev/zero to RAID50:
    10129.7218 Mbps 27 %TX 99 %RX 0 retrans 0.11 msRTT
- So, the immediate 20-Gbps challenge is primarily on the write side….



Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Nuttscp Sample Test Results Between Two "C-Systems" (3-of-7) [Source: Bill Fink/GSFC]

- Two 64-GB file copy (between four RAID5 disk controllers nested as RAID50 hosted on each C-system in a LAN testbed)
  - Configuration settings: same as in "1-of-7" **plus**
    - nuttcp application running on CPU3
  - Put file1:
    7184.8745 Mbps 41 %TX 71 %RX 0 retrans 0.11 msRTT
  - Put file2:
    7082.7940 Mbps 46 %TX 70 %RX 0 retrans 0.11 msRTT
  - Aggregate throughput:
    14267.6685 Mbps
- Better; but there was a lot of disk head contention seeking back and forth between the two files

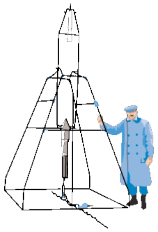
Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Nuttscp Sample Test Results Between Two "C-Systems" (4-of-7) [Source: Bill Fink/GSFC]

- A slight variation of "3-of-7", using individual 10-GigE nuttcp streams across individual 10-GigEpaths
  - Put file1:
    7136.7905 Mbps 39 %TX 72 %RX 0 retrans 0.11 msRTT
  - Put file2:
    7123.8836 Mbps 39 %TX 72 %RX 0 retrans 0.11 msRTT
  - Aggregate throughput:
    14260.6741 Mbps
- Basically the same result as "3-of-4"

Nuttscp Sample Test Results Between Two "C-Systems" (5-of-7) [Source: Bill Fink/GSFC]

- Splitting the one RAID50 into two separate RAID50s to avoid the disk head seeking contention
  - Configuration settings:
    - LRO enabled
    - eth2,3 interrupts on CPU0
    - nuttcp s2 application running on CPU1
    - 2xHPT RAID5 interrupts running on CPU2
    - first md RAID50 across above
    - 2xHPT RAID5 interrupts running on CPU2
    - second md RAID50 across above
    - nuttcp s1 application running on CPU3
  - Put file1/s1:
    9318.3251 Mbps 55 %TX 92 %RX 0 retrans 0.11 msRTT
  - Put file2/s2:
    7960.6777 Mbps 47 %TX 79 %RX 0 retrans 0.10 msRTT
  - Aggregate throughput:
    17279.0028 Mbps



Nuttscp Sample Test Results Between Two "C-Systems" (6-of-7) [Source: Bill Fink/GSFC]

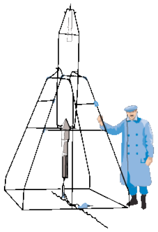
- Similar to "5-of-7" but moving the last 2 HPT RAID5 interrupts to CPU 0, so stream s2could have the same advantage as stream s1
  - Configuration settings:
    - LRO enabled
    - eth2,3 interrupts on CPU0
    - 2xHPT RAID5 interrupts running on CPU0
    - second md RAID50 across above
    - nuttcp s2 application running on CPU1
    - 2xHPT RAID5 interrupts running on CPU2
    - first md RAID50 across above
    - nuttcp s1 application running on CPU3
  - Put file1/s1:
    9161.1181 Mbps 55 %TX 94 %RX 0 retrans 0.11 msRTT
  - Put file2/s2:
    8663.7400 Mbps 52 %TX 89 %RX 0 retrans 0.11 msRTT
  - Aggregate throughput:
    17824.8581 Mbps (90% of maximum 19.8 Gbps)

Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Nuttscp Sample Test Results Between Two "C-Systems" (7-of-7) [Source: Bill Fink/GSFC]

- We are currently investigating SSD technology, to hopefully double our disk transfer speeds and get us into the 40-Gbps networked disk transfer realm

- But using parallelism of multiple cores and multiple streams is going to be key to going to 40-GigE, 100-GigE, and beyond speeds, since individual cores are not getting significantly faster



Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Nuttcp Sample Test Results With One HotLava Systems 6x10GE Tambora 120G6 NIC (1-of-8) [Source: Bill Fink/GSFC]

Configuration of Test Workstations

- Three HECN Team-assembled Intel core i7 server systems (each a 3.2-GHz single-processor (quad core) Intel Core i7 (Nehalem), over-clocked to 3.6 GHz, on an Asus P6T6 WS Revolution motherboard):
  - One using 1 HotLava 6x10-GigE NIC
  - One using 2 Myricom 2x10-GigE NICs
  - One using 1 Myricom 2x10-GigE NIC

Nuttcp Sample Test Results With One HotLava Systems 6x10GE Tambora 120G6 NIC (2-of-8) [Source: Bill Fink/GSFC]

- Theoretical maximum throughput (TMT) on a PCI-E 2.0 x16 card is nnn/(nnn+24)*8/10*16*5 Gbps, where:

nnn = PCIe MaxPayload on the test systems

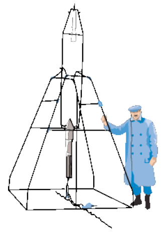
24 = PCIe protocol overhead consisting of:

1 Byte	Start of Packet (STP)
2 Bytes	Sequence Number
16 Bytes	Header (only 12 Bytes if < 4 GB address)
4 Bytes	LCRC
1 Byte	END

8/10 = 8B10B signalling encoding

16 = number of lanes

5 = 5 Gbps per lane for PCIe 2.0



Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Nuttcp Sample Test Results With One HotLava Systems 6x10GE Tambora 120G6 NIC (3-of-8) [Source: Bill Fink/GSFC]

- With MaxPayload defaulted to 128, the TMT = 53.8947 Gbps
- With MaxPayload increased* to 256**, the TMT = 58.5142 Gbps

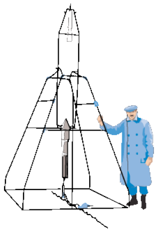
*Via the setpci command not only on the 6 10-GigE interfaces, but also on all the PCIe bridges and the Intel X58 I/O Hub in the data path, and with the PCIe MaxReadReq increased from 512 to 4096 only on the 6 10-GigE interfaces

**Not 512 because 256 is the maximum value supported by the Intel X58 I/O Hub

Nuttcp Sample Test Results With One HotLava Systems 6x10GE Tambora 120G6 NIC (4-of-8) [Source: Bill Fink/GSFC]

- 60 Second Transmit Test With MaxPayload = 128
  - n2: 8172.4651 Mbps 28 %TX 31 %RX 0 retrans 0.07 msRTT
  - n3: 8170.6930 Mbps 28 %TX 35 %RX 0 retrans 0.08 msRTT
  - n6: 8167.1622 Mbps 28 %TX 30 %RX 0 retrans 0.09 msRTT
  - n7: 8167.5251 Mbps 28 %TX 31 %RX 0 retrans 0.06 msRTT
  - n5: 8165.5400 Mbps 21 %TX 29 %RX 0 retrans 0.06 msRTT
  - n4: 8160.1735 Mbps 21 %TX 29 %RX 0 retrans 0.05 msRTT

  - Aggregate throughput:
    49003.5589 Mbps

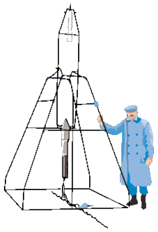


Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

Nuttcp Sample Test Results With One HotLava Systems 6x10GE Tambora 120G6 NIC (5-of-8) [Source: Bill Fink/GSFC]

- 60 Second Receive Test With MaxPayload = 128
  - n3: 9059.9860 Mbps 25 %TX 30 %RX 0 retrans 0.12 msRTT
  - n6: 8391.6758 Mbps 16 %TX 26 %RX 0 retrans 0.12 msRTT
  - n4: 8389.4628 Mbps 16 %TX 23 %RX 0 retrans 0.11 msRTT
  - n2: 9057.1408 Mbps 23 %TX 30 %RX 0 retrans 0.10 msRTT
  - n7: 8391.6331 Mbps 16 %TX 29 %RX 0 retrans 0.11 msRTT
  - n5: 8385.0556 Mbps 16 %TX 23 %RX 0 retrans 0.10 msRTT

  - Aggregate throughput:
    51674.9541 Mbps

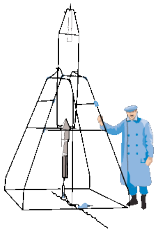


Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Nuttcp Sample Test Results With One HotLava Systems 6x10GE Tambora 120G6 NIC (6-of-8) [Source: Bill Fink/GSFC]

- 60 Second Transmit Test With MaxPayload = 256
  - n6: 9220.9229 Mbps 29 %TX 29 %RX 0 retrans 0.06 msRTT
  - n3: 9224.9003 Mbps 29 %TX 39 %RX 0 retrans 0.07 msRTT
  - n4: 9217.6819 Mbps 23 %TX 30 %RX 0 retrans 0.06 msRTT
  - n7: 9220.6031 Mbps 29 %TX 30 %RX 0 retrans 0.05 msRTT
  - n5: 9217.3856 Mbps 23 %TX 31 %RX 0 retrans 0.06 msRTT
  - n2: 9224.8250 Mbps 29 %TX 34 %RX 0 retrans 0.05 msRTT

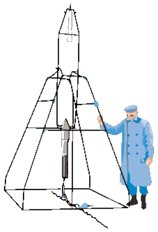
  - Aggregate throughput:
    55326.3188 Mbps (94.55 % of TMT)



Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

Nuttcp Sample Test Results With One HotLava Systems 6x10GE Tambora 120G6 NIC (7-of-8) [Source: Bill Fink/GSFC]

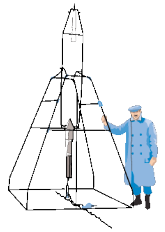
- 60 Second Receive Test With MaxPayload = 256
  - n2: 8673.6754 Mbps 22 %TX 29 %RX 1 retrans 0.12 msRTT
  - n3: 8671.5590 Mbps 24 %TX 30 %RX 0 retrans 0.10 msRTT
  - n6: 8673.8524 Mbps 16 %TX 28 %RX 0 retrans 0.12 msRTT
  - n7: 8671.3342 Mbps 16 %TX 27 %RX 0 retrans 0.10 msRTT
  - n4: 8673.6880 Mbps 17 %TX 24 %RX 0 retrans 0.10 msRTT
  - n5: 8666.1076 Mbps 16 %TX 24 %RX 0 retrans 0.11 msRTT
  - Aggregate throughput:
    52030.2166 Mbps (88.91 % of TMT)



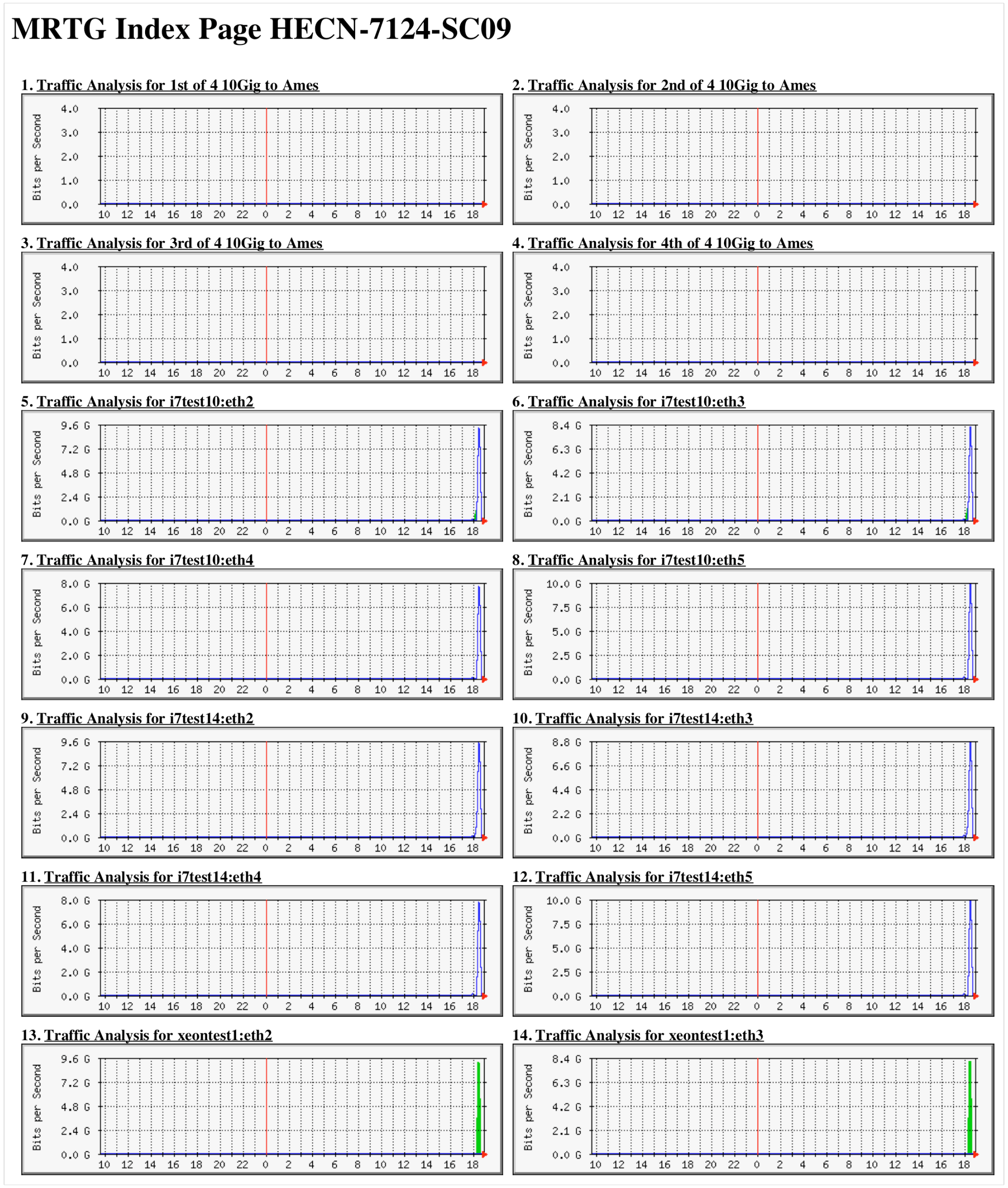
Nuttcp Sample Test Results With One HotLava Systems 6x10GE Tambora 120G6 NIC (8-of-8) [Source: Bill Fink/GSFC]

- 30 Second Bi-Directional Test With MaxPayload = 256
  - n2tx: 6834.9746 Mbps 32 %TX 48 %RX 0 retrans 0.07 msRTT
  - n6tx: 6314.6360 Mbps 33 %TX 21 %RX 0 retrans 0.19 msRTT
  - n3tx: 6195.1905 Mbps 32 %TX 38 %RX 0 retrans 0.06 msRTT
  - n4tx: 8393.6009 Mbps 28 %TX 32 %RX 0 retrans 0.05 msRTT
  - n7tx: 7489.9029 Mbps 32 %TX 27 %RX 0 retrans 0.06 msRTT
  - n7rx: 6627.6585 Mbps 11 %TX 33 %RX 0 retrans 0.23 msRTT
  - n3rx: 3264.0248 Mbps 25 %TX 33 %RX 0 retrans 2.28 msRTT
  - n2rx: 5199.5641 Mbps 37 %TX 32 %RX 0 retrans 0.10 msRTT
  - n6rx: 5117.8068 Mbps 11 %TX 33 %RX 0 retrans 0.33 msRTT
  - n4rx: 6649.6623 Mbps 15 %TX 32 %RX 0 retrans 0.27 msRTT
  - n5rx: 5815.6919 Mbps 12 %TX 34 %RX 0 retrans 6.02 msRTT
  - n5tx: 7214.4784 Mbps 32 %TX 28 %RX 0 retrans 0.07 msRTT

  - Aggregate TX throughput: 42442.7833 Mbps
  - Aggregate RX throughput: 32674.4084 Mbps
  - Total aggregate throughput: 75117.1917 Mbps



MRTG Index Page HECN-7124-SC09 10/16/09 6:59 PM



MRTG Index Page HECN-7124-SC09

1. Traffic Analysis for 1st of 4 10Gig to Ames

2. Traffic Analysis for 2nd of 4 10Gig to Ames

3. Traffic Analysis for 3rd of 4 10Gig to Ames

4. Traffic Analysis for 4th of 4 10Gig to Ames

5. Traffic Analysis for i7test10:eth2

6. Traffic Analysis for i7test10:eth3

7. Traffic Analysis for i7test10:eth4

8. Traffic Analysis for i7test10:eth5

9. Traffic Analysis for i7test14:eth2

10. Traffic Analysis for i7test14:eth3

11. Traffic Analysis for i7test14:eth4

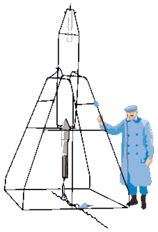
12. Traffic Analysis for i7test14:eth5

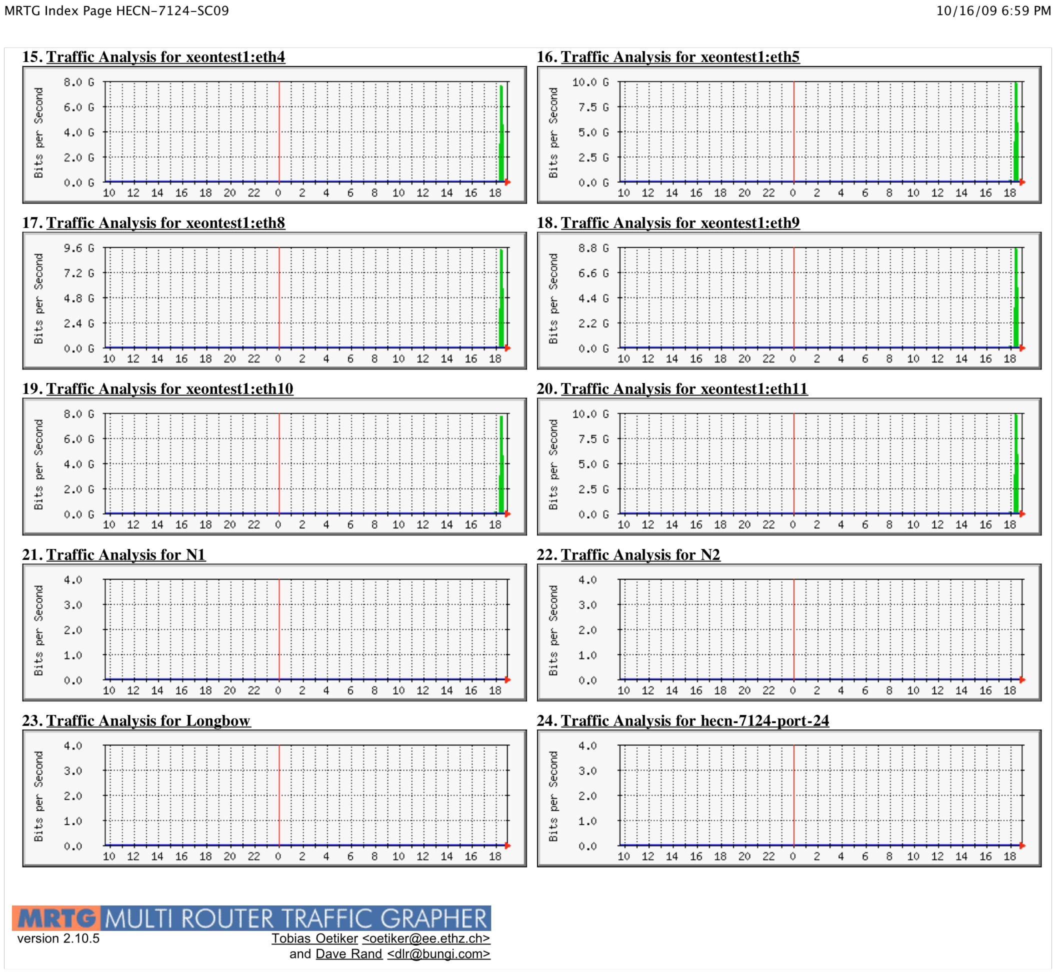
13. Traffic Analysis for xeontest1:eth2

14. Traffic Analysis for xeontest1:eth3

Bits per Second
4.0 3.0 2.0 1.0 0.0
9.6 G 7.2 G 4.8 G 2.4 G 0.0 G
8.4 G 6.3 G 4.2 G 2.1 G 0.0 G
8.0 G 6.0 G 4.0 G 2.0 G 0.0 G
10.0 G 7.5 G 5.0 G 2.5 G 0.0 G
8.8 G 6.6 G 4.4 G 2.2 G 0.0 G
10 12 14 16 18 20 22 0 2 4 6 8 10 12 14 16 18

http://shasta.sci.gsfc.nasa.gov/mrtg/gsr/hecn-7124-sc09-pat.index.html Page 1 of 2





MRTG Index Page HECN-7124-SC09 10/16/09 6:59 PM

15. Traffic Analysis for xeontest1:eth4

16. Traffic Analysis for xeontest1:eth5

17. Traffic Analysis for xeontest1:eth8

18. Traffic Analysis for xeontest1:eth9

19. Traffic Analysis for xeontest1:eth10

20. Traffic Analysis for xeontest1:eth11

21. Traffic Analysis for N1

22. Traffic Analysis for N2

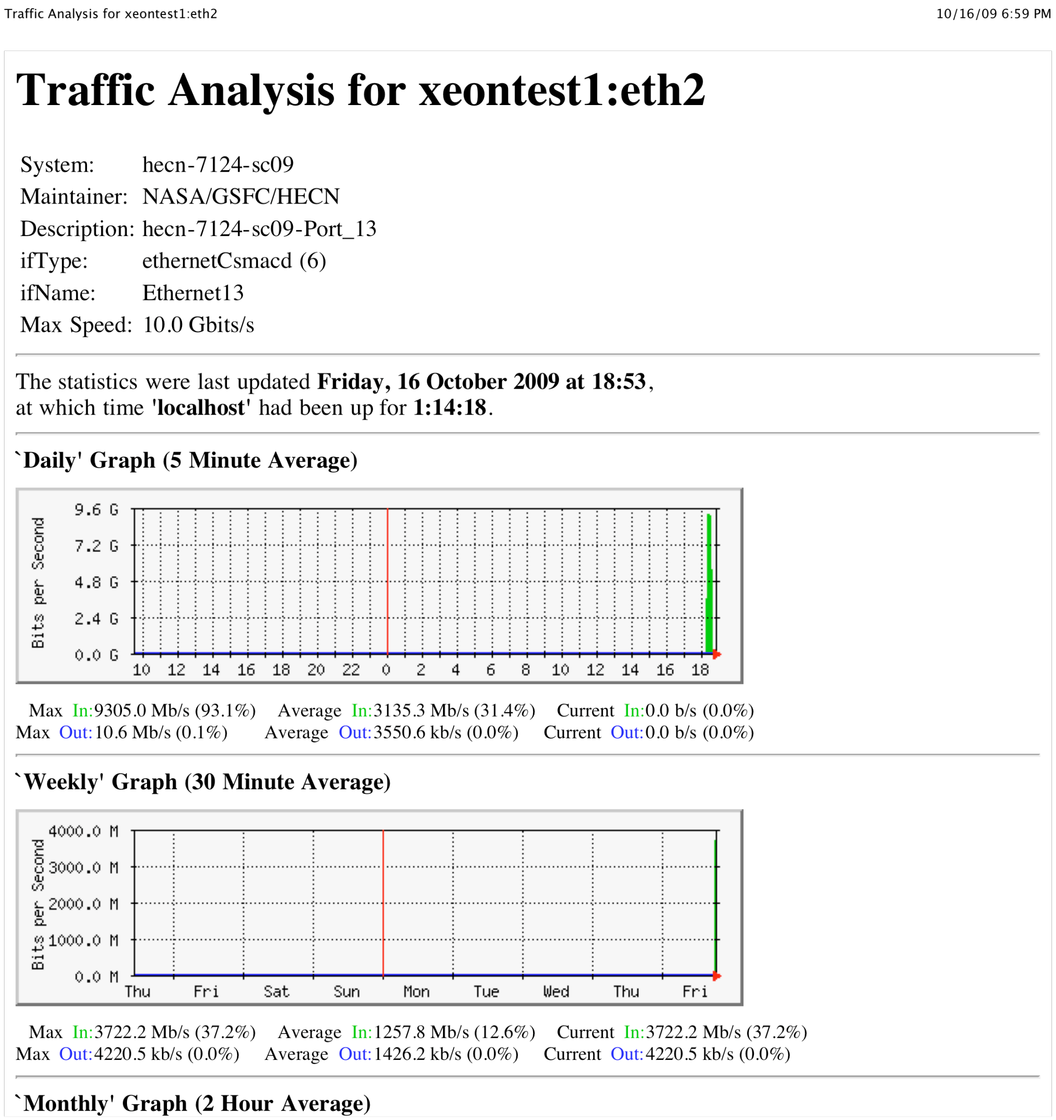
23. Traffic Analysis for Longbow

24. Traffic Analysis for hecn-7124-port-24

Bits per Second
8.0 G 6.0 G 4.0 G 2.0 G 0.0 G
10.0 G 7.5 G 5.0 G 2.5 G 0.0 G
9.6 G 7.2 G 4.8 G 2.4 G 0.0 G
8.8 G 6.6 G 4.4 G 2.2 G 0.0 G
8.0 G 6.0 G 4.0 G 2.0 G 0.0 G
10.0 G 7.5 G 5.0 G 2.5 G 0.0 G
4.0 3.0 2.0 1.0 0.0
10 12 14 16 18 20 22 0 2 4 6 8 10 12 14 16 18

MRTG MULTI ROUTER TRAFFIC GRAPHER
version 2.10.5
Tobias Oetiker <oetiker@ee.ethz.ch>
and Dave Rand <dlr@bungi.com>

http://shasta.sci.gsfc.nasa.gov/mrtg/gsr/hecn-7124-sc09-pat.index.html Page 2 of 2



Traffic Analysis for xeontest1:eth2 10/16/09 6:59 PM

Traffic Analysis for xeontest1:eth2

System: hecn-7124-sc09
Maintainer: NASA/GSFC/HECN
Description: hecn-7124-sc09-Port_13
ifType: ethernetCsmacd (6)
ifName: Ethernet13
Max Speed: 10.0 Gbits/s

The statistics were last updated **Friday, 16 October 2009 at 18:53**,
at which time **'localhost'** had been up for **1:14:18**.

`Daily' Graph (5 Minute Average)

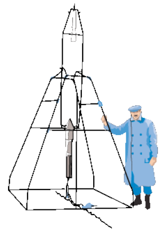
Max In:9305.0 Mb/s (93.1%) Average In:3135.3 Mb/s (31.4%) Current In:0.0 b/s (0.0%)
Max Out:10.6 Mb/s (0.1%) Average Out:3550.6 kb/s (0.0%) Current Out:0.0 b/s (0.0%)

`Weekly' Graph (30 Minute Average)

Max In:3722.2 Mb/s (37.2%) Average In:1257.8 Mb/s (12.6%) Current In:3722.2 Mb/s (37.2%)
Max Out:4220.5 kb/s (0.0%) Average Out:1426.2 kb/s (0.0%) Current Out:4220.5 kb/s (0.0%)

`Monthly' Graph (2 Hour Average)

http://shasta.sci.gsfc.nasa.gov/mrtg/gsr/hecn-7124-sc09-port_13.html Page 1 of 2

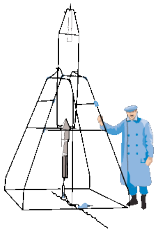


Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Phase 1 Network-Test Workstations: Nominal “B” System

- Chassis: Supermicro 836TQ-R800B (3u 16bay 7slot 800W RPS)
- Motherboard: Asus P6T6 WS Revolution (5 PCIe V2 x8)
- Processors: one Intel i7 965 (3.2GHz quad-core Nehalem)
- Memory: Kingston KHX16000D3ULT1K3 (6GB 2000MHz DDR3 CL8)
- System disks: one Western Digital WD2500BEKT (2.5” 250GB)
- NICs: two Myricom 10G-PCIE2-8B2-2S+E (Dual 10GE SFP+)
- Raid controllers: two HighPoint RocketRaid 4320 (internal, 8 disks each)
- User disks: 16 Western Digital WD5001AALS (500GB)
- IB HCA: one Qlogic QLE7280 (DDR, 8x)

- For more detail, contact Paul.Lang@nasa.gov



Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

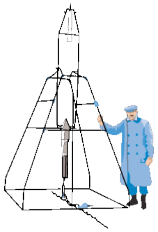
Phase 1 Network-Test Workstations: Nominal "C" System

- Nominal "B" (Baseline) System
- Minus:
  - NICs: One Myricom 10G-PCIE2-8B2-2S+E (Dual 10GE SFP+)
  - IB HCA: one Voltaire (DDR, 8x)
- Plus:
  - Raid controllers: two HighPoint RocketRaid 4322 (external, 8 disks each)
- Plus via SAS-connection:
  - Chassis: one Supermicro 836TQ-R800B (3u 16bay 7slot 800W RPS) with SAS converter/adaptor and cables
  - User disks: 16 Western Digital WD5001AALS (500GB)
- For more detail, contact Paul.Lang@nasa.gov

Phase 1 Network-Test Workstations: Nominal “A” System

- Nominal “B” (Baseline) System
- Minus:
  - Raid controllers: two HighPoint RocketRaid 4320 (internal, 8 disks each)
  - User disks: 16 Western Digital WD5001AALS (500GB)
  - IB HCA: one Voltaire (DDR, 8x)

- For more detail, contact Paul.Lang@nasa.gov

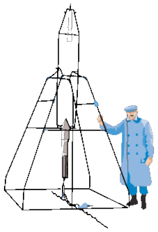


Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Phase 1 Network-Test Workstations: “A+” System

- Nominal “A” System
- Plus:
  - NICs: Two Myricom 10G-PCIE2-8B2-2S+E (Dual 10GE SFP+)

- For more detail, contact Paul.Lang@nasa.gov



Phase 1 Network-Test Workstations: “A-” System

- Nominal “A” System
- Minus:
  - NICs: One Myricom 10G-PCIE2-8B2-2S+E (Dual 10GE SFP+)

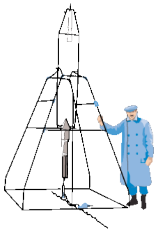
- For more detail, contact Paul.Lang@nasa.gov

Introduction To
GSFC High End Computing
20, 40 &100 Gbps Network Testbeds

Phase 1.1 Network-Test Workstations: “X++” System

- Chassis: Supermicro 836TQ-R800B (3u 16bay 7slot 800W RPS)
- Motherboard: Supermicro X8DAH+-F (6 PCIe V2 (4 x8 & 2 x16))
- Processors: two XEON W5580 (3.2GHz quad-core Nehalem)
- Memory: Kingston KHX16000D3ULT1K3 (6GB 2000MHz DDR3 CL8, running at 1333MHz)
- System disks: one Western Digital WD2500BEKT (2.5” 250GB)
- NICs: six Myricom 10G-PCIE2-8B2-2S+E (Dual 10GE SFP+)

- For more detail, contact Paul.Lang@nasa.gov



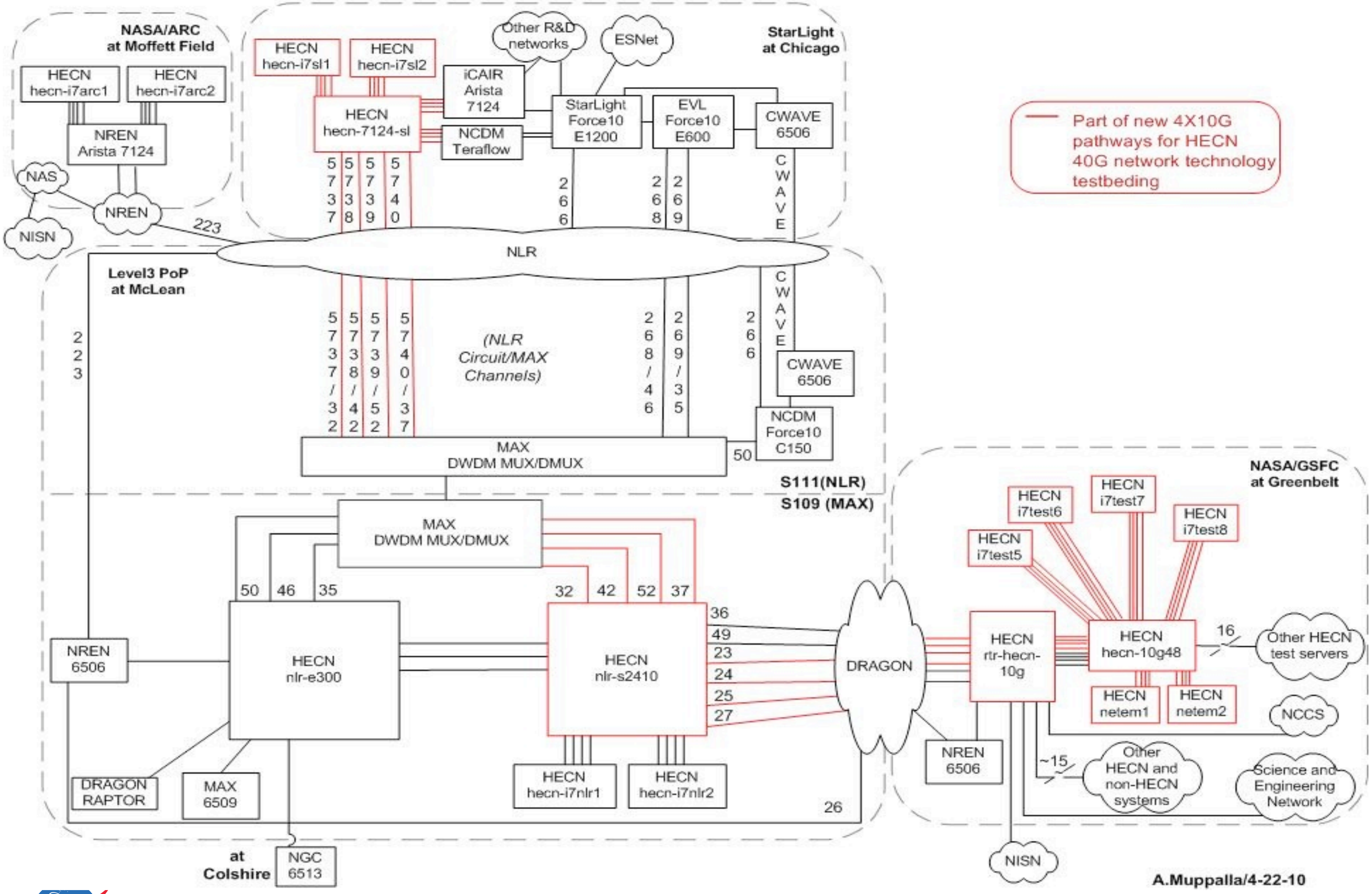
Introduction To GSFC High End Computing 20, 40 &100 Gbps Network Testbeds

“X++” Server Approximate Costs *(With components acquired via SEWP IV in lot-sizes of 3 - 15, and self assembly. Source: Paul Lang (ADNET))*

Supermicro 836TQ-R800B 3u 16bay 800W RPS Chassis	$850
Supermicro X8DAH+-F motherborad	$508
Intel W5580 XEON 3.2GHz processor $1669 x 2	$3338
Kingston KHX2000C8D3T1K3 6GB DDR3 2000 CL8 memory x 2	$500
CBL-0084 front pannel cable	$3
12" 3pin fan extension cable	$1
ArkTech slim IDE DVD to SATA adapter	$10
Myri 10G-PCIE2-8B2-2S+E Dual SFP+ NIC $950 x 6	$5700
Dynatron G666 CPU cooler	$35
Western Digital WD2500BEKT 250GB 2.5" system disk	$73
Red Greatland 18" Slimline SATA adapter	$6
Supermicro MCP-220-83601-0B FDD tray for 2.5" disk	$8
eVGA GeForce 8400GS video card	$40
8" 8pin power extension cable	$8
	$11080

GSFC/High End Computer Network (HECN) and Partners 10GE and 10G Lambda Connections Through McLean

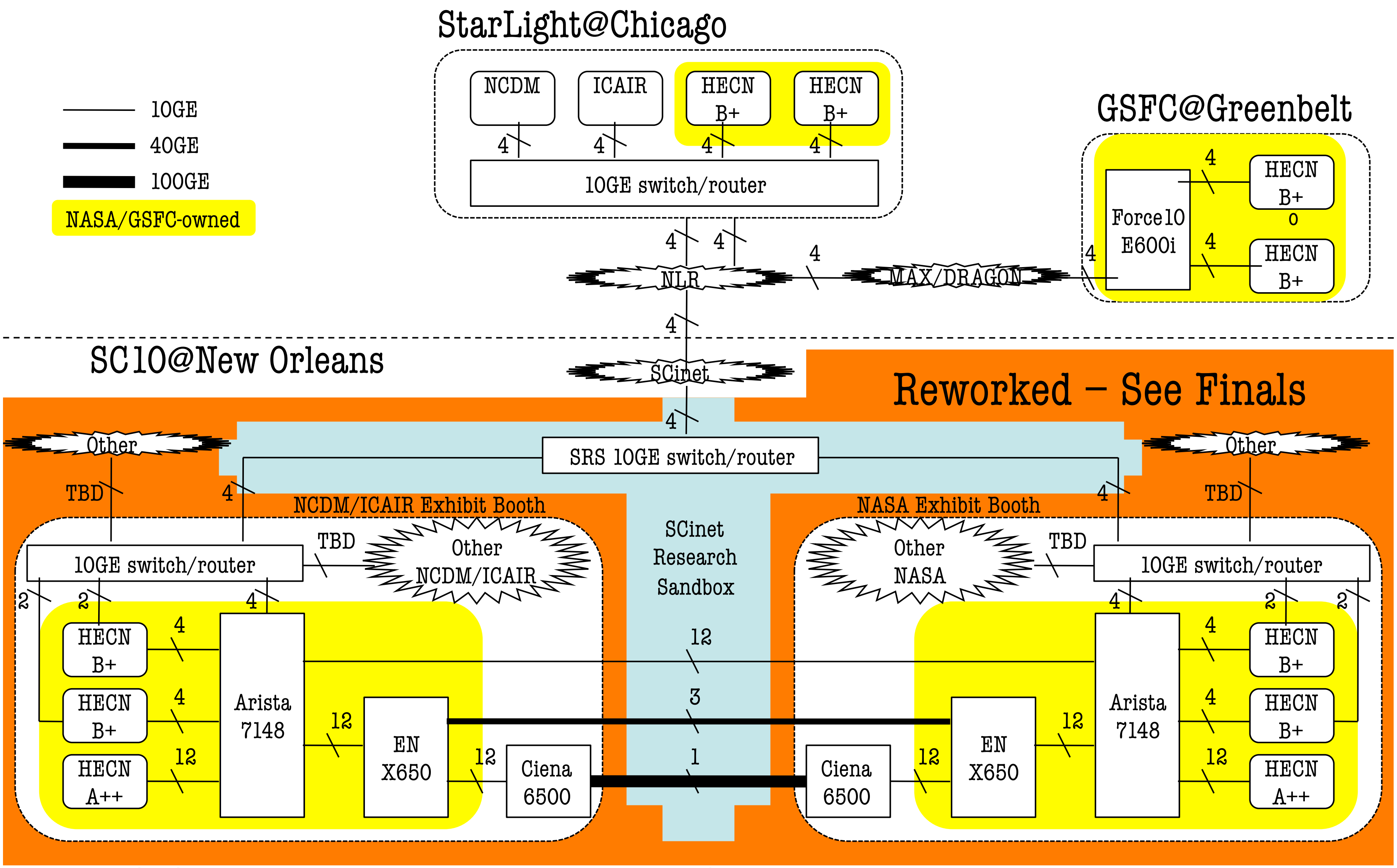
Note: The non-GSFC/HECN systems shown typically have other connections that are not shown in this diagram, as the focus is primarily GSFC/HECN connections



A.Muppalla/4-22-10

Using 100G Network Technology in Support of Petascale Science

A Collaborative Initiative Among NASA, NLR, Northwestern/iCAIR, SCinet & UIC/LAC



J. P. Gary 8/17/10

NETWORK BOTTLENECKS